# Shedding new light on random trees

Nicolas Broutin

A thesis submitted to McGill University
in partial fulfillment of the requirements of the
degree of Doctor of Philosophy

*Your file Votre référence*
*ISBN: 978-0-494-32158-4*
*Our file Notre référence*
*ISBN: 978-0-494-32158-4*

# Abstract


We introduce a weighted model of random trees and analyze the asymptotic properties of their heights. Our framework encompasses most trees of logarithmic height that were introduced in the context of the analysis of algorithms or combinatorics. This allows us to state a sort of "master theorem" for the height of random trees, that covers binary search trees (Devroye, 1986), random recursive trees (Devroye, 1987; Pittel, 1994), digital search trees (Pittel, 1985), scale-free trees (Pittel, 1994; Barabási and Albert, 1999), and all polynomial families of increasing trees (Bergeron et al., 1992; Broutin et al., 2006) among others. Other applications include the shape of skinny cells in geometric structures like $k$-d and relaxed $k$-d trees.

This new approach sheds new light on the tight relationship between data structures like trees and tries that used to be studied separately. In particular, we show that digital search trees and the tries built from sequences generated by the same memoryless source share the same stable *core*. This link between digital search trees and tries is at the heart of our analysis of heights of tries. It permits us to derive the height of several species of tries such as the trees introduced by de la Briandais (1959) and the ternary search trees of Bentley and Sedgewick (1997).

The proofs are based on the theory of large deviations. The first order terms of the asymptotic expansions of the heights are geometrically characterized using the Cramér functions appearing in estimates of the tail probabilities for sums of independent random variables.

# Résumé

Nous présentons un modèle d'arbres aléatoires pondérés et analysons les propriétés asymptotiques de leur hauteur. Notre modèle couvre la plupart des arbres de hauteur logarithmique qui apparaissent dans le contexte de l'analyse des algorithmes et en combinatoire. Ceci nous permet de formuler une sorte de "master theorem" pour la hauteur des arbres aléatoires qui recouvre les arbres binaires de recherche (Devroye, 1986), les arbres récursifs (Devroye, 1987; Pittel, 1994), les arbres digitaux de recherche (Pittel, 1985), les arbres "scale-free" (Pittel, 1994; Barabási and Albert, 1999), et toutes les familles polynomiales d'arbres croissants (Bergeron et al., 1992; Broutin et al., 2006). Certaines applications sont moins directement reliées à la hauteur des arbres. Par exemple, nous étudions la forme des cellules dans les structures de données géométriques telles que les arbres $k$-dimensionnels.

Cette nouvelle approche fait aussi la lumière sur les liens intimes qu'entretiennent les arbres et les tries, qui ont, jusqu'à présent, été étudiés de manière disjointe. En particulier, nous montrons que les arbres digitaux de recherche et les tries construits à partir de séquences générées par la même source sans mémoire partagent la même structure interne que nous appelons le "core". Ce lien entre les arbres digitaux de recherche et les tries est à l'origine de notre analyse de la hauteur des tries. Il permet, en outre, d'obtenir la hauteur des arbres introduits par de la Briandais (1959) et des arbres ternaires de recherche de Bentley and Sedgewick (1997).

Les preuves sont basées sur la théorie des grandes déviations. Le premier terme du développement asymptotique de la hauteur est caractérisé géométriquement grâce aux fonctions de Cramér intervenant dans les estimations des queues des distributions de sommes de variables aléatoires.

*do Pelagii,*

*pamiętam się i ucycę się.*

# Acknowledgement

The last three years have been wonderful. Working at McGill was like wandering on the shore to gather seashells. Sometimes, it really was. I owe a great debt to Bruce showing me that sheer force and imagination make a great team. Thanks also to Louigi, Erin, Sean, Andrew, Jamie, Vida, Conor, Babak, Perouz, and the others for making the department such a good environment.

Special thanks are due to Louiiiigi and Erin, who kept working with me despite my constant yelling: "Fucking obvious! ...Hold on ...". Someone wise said that every problem has a short, easy to understand wrong proof. After working with me, they both know that it is true, and that every problem actually has many such proofs.

Life in Montreal is definitely not dull, but it would not be the same without the lunch group. Luc, Godfried, Gabor, Vida, Perouz, Erin, and the others, shared their secrets aloud for three years, and gave me a reason to walk down to school during winter.

I think my friends in Montreal, Khrystell, Mike, Nikki, Loes and Jen, all deserve to be blessed. In particular Louigi and Rosie who supported me constantly, hosted me when I needed, and made sure I was fine at all time.

Céline was always there, and had to endure my autistic nights and week-ends. I know this was not easy, and I am glad she is still here.

I also thank my parents who kept sending me chocolate and calling me on sunday mornings when I was hangover. Thanks also to Ludovic for visiting many times and for not sleeping with anyone in my empty living room.

Finally, I am infinitely grateful to Luc. It was always challenging to follow him, whether it was standing at a blackboard, riding a bike or sitting in front of the $n$-th belgian Trappist, as $n \to \infty$. His enthusiasm and crunchy stories managed to keep me looking forward. It is funny that, in the end, I write a thesis about the first question he ever asked me, even before I had met him: "new proofs for the height of tries". This was six years ago, and back then, the height of tries was the sequence of high jumps of an athlete. I thought he wanted some kind of high precision laser measurement device, and was sure I was not his man.

# Contents

Chapter 1

# Introduction

*In this chapter, we motivate the topic of the entire document by placing the study of random trees and their heights in the necessary context. This includes, among other fields, combinatorics, computer science, and mathematical physics.*

*Auprès de mon arbre, je vivais heureux,*
*J'aurais jamais dû le quitter des yeux.*
– G. Brassens

## Contents

## 1.1 Down with determinism

No longer than eighty years ago, Einstein was, "at any rate, still convinced that He does not throw dice". It seems now that even if Einstein's God did not throw dice, he made sure that *we* would. Quantum mechanics is but one example of the ubiquity of probability in science and engineering, whether the reason be that the world is *really* random, or only that it appears so to us. From statistical physics to operations research, or economics, probability theory has proved useful in modeling, understanding, and making a better use of the world we live in. Even our everyday life is literally surrounded by chance, through its use in weather forecast for instance, with "probability of precipitation" deliberately provided to the public. Within mathematics, even in pure "deterministic" branches like number theory or geometry, major successes have been obtained using tools such as Erdős' probabilistic method (Pach and Agarwal, 1995; Alon et al., 2000). Recently, Arora and Safra (1998) proved a new characterization of the celebrated complexity class NP in terms of Probabilistically Checkable Proofs (PCP), opening a breach in deterministic complexity theory.

The increasing quantity of data involved is arguably one of the reasons of the advent of probability in today's science. It is known that the amount of data is growing at an increasing pace. DNA sequencing, high definition video, data mining are only some of the many examples illustrating this tendency. Believing Spinoza's maxim that "nature abhors a vacuum", Kryder's law, the storage analog of the well-known Moore's law, surely accounts for this fact. The now phenomenal volume of information makes it not only relevant, but also necessary to step back in order to look at data at a larger scale: finding a macroscopic structure in the microscopic (apparent?) chaos. Large quantities justify a statistical approach that focus on obtaining a glance of the big picture. This observation alone justifies the use of probability together with its modern machinery.

## 1.2 Randomness and computing

Historically, both computational problems and the algorithms solving them have been studied for the worst case input. In other words, one wanted to make sure that a problem admitted a method to find the solution, or that an algorithm was efficient for *any* possible input. Then, it is no surprise that the complexity classes P and NP were the first to be introduced (see Garey and Johnson, 1979). However, when studying algorithms, one quickly notices that they behave usually very nicely for "standard" inputs, and may fail or run longer on some particularly nasty inputs —in particular, the ones used to build *gadgets* and prove hardness. Although there is great value in worst-case analysis, since these extreme cases *can* happen, there is probably as much value in an analysis telling *quantitatively* what *will likely* happen and what *should not* happen in general. One can make the notion of "standard input" (a bit) more precise by agreeing that most inputs should be standard. One can require that an algorithm (Monte-Carlo) should be efficient on all inputs, and correct on *most inputs*: this gives rise to the class BPP (bounded error, probabilistic, polynomial time). Or, one can require that it be correct on most inputs, and efficient *on average* (Las Vegas): this is RP (randomized, polynomial time), and ZPP (zero-error, polynomial time) (see, e.g., Sipser, 2005; Papadamitriou, 1993). In this latter case, an algorithm is also allowed to work very slowly (non-polynomially) on some inputs, but it should be polynomial on average. Such an approach is also justified by the fact that many algorithms are rarely used on a single input: in the long term, that is, if one runs the algorithm a large number of times on *different* inputs, one may expect that the nasty cases occur very rarely, and thus that *on average* the algorithm runs much faster than on the worst possible input.

In the early seventies, under the impulse of Knuth (1973a,b,c), a community of mathematicians and computer scientists took this point of view and started to analyze efficient algorithms *on average.* They were pioneers of what is now known as the *analysis of algorithms.* Of course, one needs first to define what they mean by *on average*, and thus to define the model of randomness. However, it is usually natural, and unless one expects some particular pattern in the input (because it is generated by some other algorithm, for instance) it may be considered relatively *random* (under a model of randomness to be specified). If it is not the case, one can often cope with the non-randomness by preprocessing the input to make it random at little cost (applying some sort of permutation, for example).

One of the most celebrated results of the analysis of algorithms concerns Hoare's Quicksort algorithm (Hoare, 1961, 1962), which recursively sorts a collection of $n$ entries. It is known to run in time $\Theta(n^2)$ in the worst case. However, if one either randomizes the algorithm (or, equivalently, permutes randomly the input), then it runs in time $O(n \log n)$ on average, which makes it optimal within a multiplicative constant (Sedgewick, 1975; Sedgewick and Flajolet, 1996). The particular appeal of the questions arising and the connections with other fields of mathematics (like, for instance, complex analysis or information theory) attracted many researchers in the footsteps of the pioneers of analysis of algorithms. Researchers now succeed in deriving detailed information about, among other things, the running times or storage space required by most important algorithms: not only the mean, variance and other moments, but also limit distributions and tail probabilities are sought after (Vitter and Flajolet, 1990; Devroye, 1998a). A deep understanding of the phenomena underlying the behavior of algorithms made it possible to *design* efficient algorithms that take advantage of these observations (e.g., Flajolet, 2004).

## 1.3 Of the importance of trees

There is no doubt that trees are ubiquitous when dealing with algorithms, either for storing, manipulating or even representing data (Cormen et al., 2001). Whether the data should be organized as a priority queue (heaps, Fredman and Tarjan (1987)), a dictionary (search trees, Sleator and Tarjan (1985)), a collection of mergeable sets (link-cut trees, Tarjan (1983)), or a compact representation of proximity (minimum spanning trees or Steiner trees, Barthélemy and Guénoche (1991)) the most efficient structures are often based on trees. Also, the branching structure of algorithms is arguably one of the main reasons for studying trees to understand how algorithms behave.

Even outside of the field of computing sciences, trees are of prime importance. Computer science is still young, and it is not surprising that branching structures have appeared and been studied first in fields that are as various as biology, physics as well as social sciences. Some of the most surprising examples of the use of trees include literature, politics and scotch tasting! We shall only give a few examples that should convince the reader of the wide range of applications of tree structures. Probably one of the earliest applications of trees to a concrete problem is that of Galton (1873) and his famous study of the pool of family names in England. We shall look at this example more carefully later in Chapter 3. In biology, phylogenetic trees have been used to study the spread of epidemics and the evolution of species (Barthélemy and Guénoche, 1991). Phylogenetic trees have also been used in literature to study formally the work of authors such as Shakespeare and Giraudoux (Barthélemy and Luong, 1987). Such analysis reveals not only the usually accepted classifications but open also new directions of investigation.

As a consequence, judging from the ubiquity of trees in all the fields of research, there is no doubt that a better understanding of tree structures will *some day* be of some interest to a researcher or another, whether he be a mathematician, a physicist, a linguist or just happened to be caught be the beauty of trees.

## 1.4 Random trees and their heights

### 1.4.1 A model of randomness

We claimed that useful information can be derived from the analysis of random version of trees, but we have not yet told anything about the model of randomness. There are many natural models of interest. For instance, one could consider a tree taken uniformly at random in the set of trees of a certain class (binary trees, rooted trees, rooted plane trees, etc.). This approach has been taken by Flajolet and Odlyzko (1982) who managed to capture asymptotics of the heights of a large class of trees using a single method. Such random trees have typical heights and widths of order $\Theta(\sqrt{n})$. The trees arising under this model of randomness appear mostly in combinatorics (Cayley trees, simply generated trees), probability theory (Galton–Watson trees conditioned on the size), or statistical physics. They have been cast away from computing applications, for a good reason: their lengthy branches make them usually inefficient when one has to traverse the structure.

A model of randomness that is sometimes more pertinent is to construct the tree sequentially from random inputs (we will explain what we mean by this shortly), and then study the tree obtained. The trees built by such a procedure are typically of logarithmic height. So, they are bushier and more compact that the uniform trees of the previous paragraph. This is why they often appear when analyzing efficient data structures and algorithms. One should also observe that a randomization of the inputs seems more natural when dealing with algorithms than a randomization of the trees themselves. We are interested in this latter class of models. Our objective is to devise a unifying approach for analyzing the heights of such trees. In this sense, our project can be seen as complementing that of Flajolet and Odlyzko (1982) for uniform trees. In what follows, we consider those random trees with a logarithmic height only.

### 1.4.2 A canonical example

We now introduce the main concepts we will deal with using a celebrated example, namely binary search trees (BST). Consider a set of $n$ distinct keys $\{x_1, x_2, \ldots, x_n\}$ that one wants to store in a tree to handle search queries. The keys need to be comparable, and without loss of generality (use their ranks), we can assume they are elements of $\{1, 2, \ldots, n\}$. The binary search tree associated with the sequence $\{x_1, \ldots, x_n\}$ consists of a node storing the first key $x_1$, and of two subtrees. The left and right subtrees are the binary search trees (recursively built) associated with the sequences $\{x_i : x_i < x_1\}$ and $\{x_i : x_i > x_1\}$, respectively. Figure 1.1 shows a binary search tree on $\{1, 2, \ldots, 9\}$. If the input is a uniform random permutation $\{X_1, \ldots, X_n\}$ of $\{1, 2, \ldots, n\}$, the tree is called a *random binary search tree.*

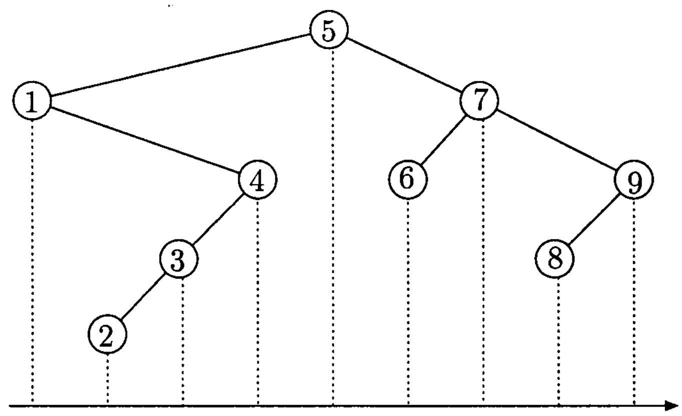


***Figure 1.1:*** *A binary search tree of build on the sequence* $\{1, 2, \ldots, 9\}$.

One can construct a random binary search tree incrementally instead of recursively. This is done by assigning the keys $x_i$ to the nodes of a complete infinite binary tree $T_\infty$. The first tree $T_1$ consists of a single node storing $x_1$. Let $T_i$ be the tree built from the sequence of keys $\{x_1, \ldots, x_i\}$. Let $i \geq 2$ and assume $T_{i-1}$ has already been built. To insert $x_i$, we start from the root and go down the tree by moving left at a node if the key it stores is greater than $x_i$, and right otherwise. We place $x_i$ in the first empty node we find. The first node on each path down the root is called *external.* Note that at each step, the next node $x_i$ is placed in a node of the *fringe* of $T_i$, the set of external nodes (see Fig. 1.2). The resulting random binary search tree consists precisely of those nodes that have been assigned a key.

In spite of their remarkably simple definition, random binary search trees are at the heart of analysis of algorithms. It is interesting to note that random binary search trees are distributed like the branching structure of a randomized version of Quicksort. Another reason for their success is the fact that they concentrate many of

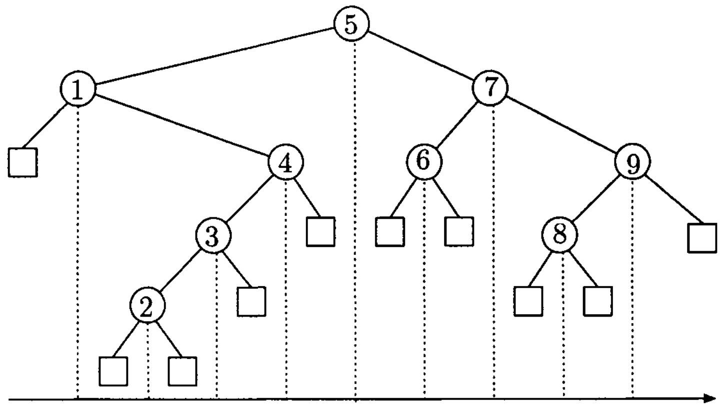


*Figure 1.2: The binary search tree of Figure 1.1 with its fringe (also called the external nodes) represented with square boxes.*

the central questions of analysis of algorithms in one of the simplest, yet interesting models. Let us illustrate this fact by going back the the incremental growing process. The time to build the tree is (the number of nodes plus) the sum of the depths of the nodes, or the *path length.* The height of the tree (the largest depth of a node) is the maximum time it takes to insert a node. Equivalently, if one then uses the tree for queries, the average depth of a node in the tree is the average cost of a *successful* search; the height is the maximum cost of such a search. So the height of the random tree corresponds to the worst case cost in some average sense. It happens that, in this average sense, the worst case time is still $O(\log n)$ and it is much better than the deterministic worst case of $\Theta(n)$.

### 1.4.3 The height of binary search trees

Pinning down the height $H_n$ of a random binary search tree of size $n$ was one of the central problems in analysis of algorithms. The hunt for the asymptotic properties of $H_n$ has tied many profound bonds between the analysis of algorithms, statistical physics, and the theory of branching processes. Robson (1979) was the first to prove the upper bound: for $c = 4.311\ldots$, the unique solution greater than 1 of $c\log(2e/c) = 1$, for any $\epsilon > 0$, $\mathbf{P}\{H_n \geq (c+\epsilon)\log n\} \to 0$ as $n \to \infty$. Robson (1982) then realized that $\mathbf{E}H_n \sim \gamma \log n$ as $n \to \infty$, and his experimental simulations seemed to indicate that $c$ was the best candidate for the limit constant.

One of the crucial ideas is due to Pittel (1984) who provided the connection between the height of binary search trees and branching random walks that paved the road to the solution. Pittel's idea is better explained using the incremental construction of random binary search trees we have described in Section 1.4.2. It is based on the following observation: since $\{X_1, X_2, \dots, X_n\}$ is a random permutation of $\{1, 2, \dots, n\}$, the rank (position in the ordered list) of $X_n$ is uniform in $\{1, \dots, n\}$. Therefore, $X_n$ is equaly likely to be stored in any of the $n$ external nodes of $T_{n-1}$, the tree built from $\{X_1, \dots, X_{n-1}\}$. This evolution process can be simulated using random clocks. Given $T_{n-1}$ with the right distribution, if the edges leading to the external nodes have random clocks, and any of them is equaly likely to tick first, then putting $X_n$ at the corresponding position yields a tree distributed as $T_n$. Adding $X_n$ creates two new external nodes that, in turn, would be equaly likely to be picked thanks to similar random clocks.

We have to make sure that, at any stage of the process, the next clock to tick is uniformly random, and *independent* of the time when the clock came into play. There is a well-know way to achieve this, which uses exponential random variables and their memoryless property. Let $\{T(t), t \geq 0\}$ be the continuous-time branching process that will eventually be used to embed $T_n$. At time $\tau_1 = 1$, $T(\tau_1)$ consists of a single node, together with two independent clocks on the edges to the fringe. One of the clocks ticks at time $\tau_2$ and gives birth to a node $u_2$. Any clock stops once it has ticked. Assume now that at time $t$, the tree consists of $n-1$ nodes and $n$ clocks (on the edges to potential future nodes). A uniformly random clock ticks at time $\tau_n$, giving birth to $u_n$ and its two new clocks. Then, for every $n \geq 1$, $T(\tau_n)$ is distributed as $T_n$. Also, the height of $T_n$ is the maximum number of edges on a path down the root, and this corresponds to the maximum number of ticks that occured on the same line of descent before $\tau_n$.

The process described above may be seen as first-passage percolation on $T_\infty$. First put down all the exponential clocks we may ever need: Assign independent and identically distributed (i.i.d.) exponential(1) random variables to the edges, say $\{E_e, e \in T_\infty\}$. Let $\pi(u)$ denote the set of edges on the unique path from $u$ to the root. A node $u$ is born at time $B_u = \sum_{e \in \pi(u)} E_e$. For $t \geq 0$, let $T(t)$ be the subtree of $T_\infty$ consisting of the nodes born before time $t$:

$$T(t) = \{u \in T_\infty : B_u \leq t\}.$$

Then, at the random time $\tau_n = \inf\{t : |T(t)| \geq n\}$, the tree $T(\tau_n)$ is distributed as $T_n$ with probability one. By tweaking the model, and introducing a little dependence in $\{E_e, e \in T_\infty\}$, one can make $\tau_n \leq \log n < \tau_{n+1}$ deterministically, hence yielding the property that $T(\log n) = T_n$ in distribution (Devroye, 1986).

Then, asking for the height of $H_n$ reduces to finding $k$ such that there is a node in the $k$-th generation that is born before time $\log n$, but none in the $(k+1)$-st one. Or, turning the question upside down, one only needs to characterize the random time of the first birth of a node at level $k$. Using his continuous embedding and subadditive arguments, Pittel proved that there exists $\gamma > 0$ such that $H_n \sim \gamma \log n$ almost surely, as $n \to \infty$. Taking advantage of the Hammersley–Kingman–Biggins theorem (Hammersley, 1974; Kingman, 1975; Biggins, 1977) about the first-birth problem in branching random walks, Devroye (1986) finally showed that Robson's upper bound was indeed tight.

**Theorem 1.1 (Devroye 1986).** *Let $H_n$ be the height of a random binary search tree of size $n$. Then, $H_n \sim c \log n$ in probability, where $c$ is the unique solution greater than 1 of $c \log(2e/c) = 1$.*

Theorem 1.1 is far from being the end of the story. Robson (1982) noticed that, when simulated experimentally, $H_n$ exhibited very little variance and conjectured that it has bounded variance, $\mathbf{Var} H_n = O(1)$. Robson's variance conjecture became the next hot topic. Pushing his branching processes techniques further, Devroye (1987) proved that $\mathbf{Var} H_n = O(\sqrt{\log n \log \log n})$. The bound was later improved by Devroye and Reed (1995) who showed using the second moment method that $\mathbf{Var}H_n = O(\sqrt{\log\log n})$. In the mean time, Drmota (2001) derived an alternative proof of Theorem 1.1 using analytic tools and generating functions, giving new credit to Robson's conjecture: using his novel point of view, he observed that all central moments of $H_n$ might be bounded as well. Finally, Robson's conjecture was proved by Reed (2000), Reed (2003) and Drmota (2003) and it is now known that $\mathbf{Var}H_n = O(1)$ and

$$\mathbf{E}H_n = c\log n - \frac{3}{2\log(c/2)}\log\log n + O(1).$$

### 1.4.4 Towards unification

The first moment of the height of random binary search trees remained a question of interest. For instance, slight modifications in the proof of Devroye (1986, 1987) proved successful in obtaining asymptotic properties of other random trees like random recursive tres (Devroye, 1987; Pittel, 1994), $m$-ary search trees (Devroye, 1990; Pittel, 1994), pyramids (Mahmoud, 1994; Biggins and Grey, 1997). It seemed apparent that the branching processes arguments were suitable to unify all these scattered results, see Devroye (1987), Pittel (1994), Biggins and Grey (1997) and Devroye (1998b).

It is interesting to note that researchers also worked at generalizing the theorems about higher moments as well. In particular Chauvin and Drmota (2007) proved, as in the binary case, the height of $m$-ary search trees has the distribution of a travelling wave. This is closely related to a finer analysis of the first-birth problem and the work of Bramson (1978) and Bachmann (2000). Drmota (2006) has proved that increasing trees (a class of random trees that encompasses binary search trees, and random recursive trees) exhibit a similar behavior, although the average height is not characterized. Very recently, using a deep connection with generalized ballot theorems, Addario-Berry and Reed (2006) proved that for a large class of branching random walks, the average height is of the form $\alpha\log n - \beta\log\log n + O(1)$ (see Addario-Berry, 2006; Addario-Berry and Reed, 2007). Results on higher moments follow easily from the precise position of the mean.

Although there are beautiful connections arising in the analysis of the distribution of the heights of random trees, there is still much to be done about the average value. The kind of generalization we are after is one that would allow a large class of trees to be dealt with a unique framework.

### 1.4.5 What about tries?

While tries are part of the class of random trees of logarithmic height, they have usually been studied separately. We try here to (1) exhibit the very reason why they were put aside, and (2) reconcile them with the rest of the class of trees of logarithmic height. But we shall first introduce them and review the previous work on the heights of tries.

Tries are data structures used to manipulate and store strings by taking advantage of the digital character of words. They were introduced by de la Briandais (1959). Apparently, the term of *trie* was coined by Fredkin (1960) as a part of the word "retrieval". Their properties and uses are reviewed by Knuth (1973c) and more recently by Szpankowski (2001). Consider $n$ sequences of characters (or strings) from a countable alphabet $\mathcal{A}$. Each one of the sequences carves an infinite path in an infinite rooted position tree $T_\infty$ where the children of each node are labeled with the characters of $\mathcal{A}$: starting from the root, the characters are used in order to move down the tree. If all the sequences are distinct, the corresponding paths in $T_\infty$ are distinct as well. The trie $T_n$ is defined to be the smallest subtree of $T_\infty$ such that the paths corresponding to the sequences are distinct within $T_n$. A trie on the binary alphabet $\{0, 1\}$ is shown on Figure 1.3. When the sequences are the successive suffixes of the same string, the trie is called a suffix tree. Suffix tree are particularly important in lossless data compression like Lempel–Ziv algorithms (see Ziv and Lempel, 1977, 1978), and their tight relationship with tries build from independent sequences should suffice to motivate the study of the latter trees.

Random tries can be built by using random sequences as an input. Many models

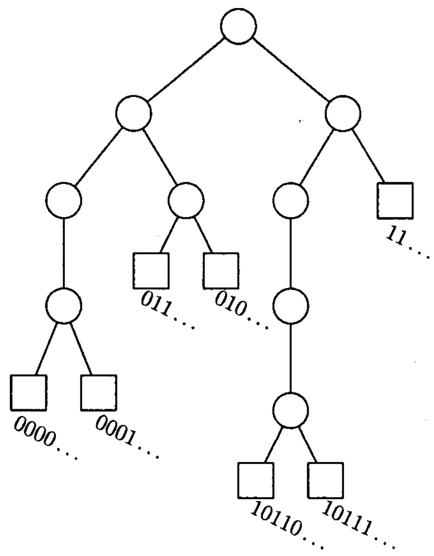


*Figure 1.3:* *The trie built from the words* 0000..., 0001..., 11..., 10110..., 011..., 010..., *and* 10111....

of randomness have been considered to fit with the statistical properties of the data used in practice. We can cite, in particular, memoryless sources (independent coin flips), Markovian sources (the successive states of a Markov chain), the so-called density model, or even continued fraction expansions and dynamical sources. For more information, we refer the reader to the textbook of Szpankowski (2001) and the recent survey by Flajolet (2006).

We now consider only the tries built using memoryless sources, that is when the sequences are independent sequences of i.i.d. coin flips. The average height has been studied by Devroye (1984), Pittel (1985) and Szpankowski (1991) (under a slightly more general model). Results on the limit distributions can be found in Flajolet (1983), Jacquet and Régnier (1986), and Pittel (1986). Devroye (2002, 2005) analyzes the concentration properties of many parameters of tries and gives strong tail inequalities. We aim at some kind of generalization of these results. Recently, Park, Hwang, Nicodème, and Szpankowski (2006) have unified the analysis of many parameters related to tries via the number of nodes at each level, also called the *profile* (see also Hwang, 2006). On the other side of the spectrum, the generalization provided by Clément et al. (1998, 2001) deals with very general sources, as well as many different trie structures, including the trees of de la Briandais (1959) and the ternary search trees (TST) of Bentley and Sedgewick (1997). The question of height of these two latter structures has been left open. In any case, the proofs are based on an analysis of the longest common prefix of a pair of sequences. So, the techniques are based on *words* and information theory rather than the trees themselves and it is not surprising to observe that the theory of branching processes appears useless in this case.

As for the case of trees based on branching processes, our goal is to devise a framework that would encompass a large class of trie structures. In particular, our class should cover the trees of de la Briandais (1959) and the TST.

## 1.5 Thesis contributions

In this document, we develop a general framework to analyze heights of trees. We distinguish two classes of trees: whether the height is bounded or not. We say that $T_n$ has *bounded height* if there exists a *deterministic* function $\psi$ such that the height $H_n \leq \psi(n)$. Tries do not have bounded height since the trie built from two identical sequences is an infinite path. We shall refer to the class of branching structures with bounded height as *trees* as opposed to *tries* (although, sensu stricto, tries are trees as well). The *bounded-height* property is not a mere remark, and we think it is the main reason that has prevented researchers from unifying the heights of tries and other trees of logarithmic heights. In some sense, we distinguish trees from tries, because they ought to be distinguished. However, we aspire at unifying both parts, and the glue should consist in a mix of profiles and large deviation theory. In the entire document, we always aim at emphasizing the geometric representations of the phenomena, as well as making explicit the underlying intuition.

CHAPTER 5: WEIGHTED HEIGHT OF RANDOM TREES. We start by introducing a model of *ideal trees*. This is just a slight generalization of branching random walks. Analyzing the height of ideal trees is directly related to the first-birth problem. We then rely on the intuition given by this idealized model to develop a general model of weighted random trees. The model allows to capture the properties of the heights of many known random trees with *bounded heights* including binary search trees, median-of-(2k+1) trees, random recursive trees, scale-free trees, random increasing trees, digital search trees, and a pebbled version of ternary search trees and list-tries. This first part is based on joint papers with L. Devroye, E. McLeish and M. de la Salle.

CHAPTER 6: WEIGHTED HEIGHT OF RANDOM TRIES. We then extend the results for certain classes of weighted tries. The ideas are based on the study of the profiles of these trees. The profile is indeed the connection between the tree of bounded heights like digital search trees and PATRICIA, and their trie counterparts. In particular, if the sequences are generated using the same source, the profile of the trie is a slight modification of that of the corresponding digital search tree. The modification is very simply described geometrically. This allows us to capture asymptotic properties of the heights of the trees of de la Briandais (list-trie) and of ternary search trees (BST-trie). This part is joint work with L. Devroye.

Chapter 2

# Probability and Large Deviations

*We recall here the probabilistic background necessary to understand this thesis. For a comprehensive account of probability theory, see Grimmett and Stirzaker (2001). For a measure theoretic point of view we refer the reader to Billingsley (1995). The two volumes of the treatise of Feller (1968, 1971) are also wonderful references. It turns out that the first order asymptotics of the heights of random trees with logarithmic heights is tightly captured by large deviations for sums of random variables. We start by giving some intuition using basic probabilistic tools. We then present in more detail the theorems of Cramér and Gärtner–Ellis. Finally, we give some useful properties of the rate functions involved. For a more complete treatment of large deviations and its applications, see Ellis (1985), Deuschel and Stroock (1989), Dembo and Zeitouni (1998), or den Hollander (2000).*

*Alea jacta est.*
– J. Cesar.

## Contents

## 2.1 Generalities

### 2.1.1 Basic notations

Throughout the document, we let $\mathbb{R}$ and $\mathbb{N}$ denote the set of real and natural numbers, respectively. We use log for the natural logarithm in place of ln; for $a > 0$, $\log_a$ stands for the logarithm in base $a$. For nonnegative sequences $x_n$ and $y_n$, we describe their relative order of magnitude using Landau's $o(\cdot)$ and $O(\cdot)$ notation. We write $x_n = O(y_n)$ if there exist $N \in \mathbb{N}$ and $C > 0$ such that $x_n \leq Cy_n$ for all $n \geq N$. Occasionally, we write $x_n = \Omega(y_n)$ to mean that there exists $N \geq 0$ and $C > 0$ such that for all $n \geq N$, $x_n \geq Cy_n$. If $x_n = O(y_n)$ and $x_n = \Omega(y_n)$, the we write $x_n = \Theta(y_n)$. If $x_n$ converges to $x$ as $n$ goes to infinity, then we write $x_n \to x$, as $n \to \infty$. An equivalent notation for $x_n/y_n \to 0$ as $n \to \infty$ is $x_n = o(y_n)$. We write $x_n \sim y_n$ when $x_n/y_n \to 1$ as $n \to \infty$.

For a function $f$, we write $\mathcal{D}_f$ for its domain $\{x : |f(x)| < \infty\}$. The interior of a set $\Gamma$ is denoted by $\Gamma^o$. The derivative of $f$ at a point $x_o$ is denoted

$$\left.\frac{\partial f(x)}{\partial x}\right|_{x=x_o}.$$

### 2.1.2 Probability

We let $\mathbf{P}\{A\}$ denote the probability of an event $A$, i.e., a measurable set defined on some probability space. We usually do not make explicit reference to the probability space since it is usually clear to which one we are referring. We say that an event $A$ holds *almost surely* (a.s.) if $\mathbf{P}\{A\} = 1$. The random variables considered in this document take values in $\mathbb{R}$ or $\mathbb{R}^d$ for some $d \in \mathbb{N}$. The expected value of a real valued random variable $X$ is denoted by $\mathbf{E}X$ or $\mathbf{E}[X]$, its variance by $\mathbf{Var}[X]$. The expected value of $X$ conditional on $A$ is written $\mathbf{E}[\,X \mid A\,]$. The indicator function of an event $A$ is of particular interest, it is denoted by $\mathbf{1}[A]$ and we have $\mathbf{E}\mathbf{1}[A] = \mathbf{P}\{A\}$.

Consider a sequence of random variables $X_n, n \geq 0$. We say that $X_n$ converges to $X$ *in probability* if for any $\epsilon > 0$, $\mathbf{P}\{|X_n - X| \geq \epsilon\} \longrightarrow 0$, as $n \longrightarrow \infty$. We write $X_n \sim a_n$ in probability if $X_n/a_n \longrightarrow 1$ in probability, as $n \longrightarrow \infty$. The sequence $X_n$ converges to $X$ *almost surely* if $\mathbf{P}\{\lim_{n\to\infty} X_n = X\} = 1$. We say that $X_n$ converges to $X$ *in distribution*, and we write $X_n \xrightarrow{d} X$ if $\mathbf{P}\{X_n \leq x\} \longrightarrow \mathbf{P}\{X \leq x\}$, as $n \longrightarrow \infty$, for all points of continuity of the *distribution function* $F : x \mapsto \mathbf{P}\{X \leq x\}$.

## 2.2 Rare events and tail probabilities

The classical limit theorems in probability theory deal with sums of independent and identically distributed (i.i.d.) random variables, so it is not a bad idea to introduce our concepts using these well known settings. Our presentation is largely inspired of the insightful introduction of Dembo and Zeitouni (1998). Consider a sequence of i.i.d. random variables $\{X_i, i \geq 1\}$. Let $S_n$ denote the sequence of their partial sums. So

$$S_n = \sum_{i=1}^{n} X_i.$$

If $\mathbf{E}|X_1| < \infty$, then Kolmogorov's strong law of large numbers asserts that $S_n/n \longrightarrow \mathbf{E}X_1$ almost surely, as $n \longrightarrow \infty$. So we expect $S_n/n$ to be close to $\mathbf{E}X_1$. The next natural interesting question arising is: "*how close?*". A first simple answer is given by Chebychev's inequality: for any $t > 0$,

$$\mathbf{P}\{S_n - n\mathbf{E}X_1 \geq nt\} \leq \frac{\mathbf{Var}[\,X_1\,]}{nt}. \tag{2.1}$$

As a consequence, if $\mathbf{Var}[\,X_1\,] < \infty$, the probability that $S_n$ exceeds its mean by a linear amount decays at least polynomially in the number of variables. The tail probability in the left-hand side of (2.1) is called a large deviation tail probability. The bound given in (2.1) is usually far from tight. To understand why, assume that $X_1, \ldots, X_n$ are independent Gaussian random variables with mean zero and unit variance. Then, $S_n$ is also a Gaussian random variable, with mean zero and variance $n$. In other words, $S_n/\sqrt{n}$ is a standard Gaussian random variable, and for all $t \in \mathbb{R}$,

$$\mathbf{P}\{S_n \geq nt\} = \mathbf{P}\left\{\frac{S_n}{\sqrt{n}} \geq t\sqrt{n}\right\} = \frac{1}{\sqrt{2\pi}} \int_{t\sqrt{n}}^{\infty} e^{-x^2/2} dx.$$

It is now clear from (2.2) that, as $n \to \infty$,

$$\mathbf{P}\{S_n \geq nt\} \;=\; e^{-nt^2/2+o(n)}.$$

So, under our assumptions, we expect that the large deviation tails for sums of i.i.d. Gaussian random variables be exponential in the number of variables. This is the kind of tail bound we are interested in because they are the ones that are relevant when studying the heights of random trees. It can be proved in a far more general setting that such exponential tail bounds hold, and are tight.

## 2.3 Cramér's Theorem

Although Cramér's theorem (Cramér, 1938; Chernoff, 1952) is the easiest of the theorems dealing with large deviations, it is still a powerful tool. We consider a sequence of i.i.d. random variables $\{X_i, i \geq 1\}$ distributed like $X$, taking values in $\mathbb{R}$. Write $S_n = \sum_{i=1}^n X_i$. We are interested in proving exponential bounds for the (right) tail probability $\mathbf{P}\{S_n \geq tn\}$ when $t \geq \mathbf{E}X$, as $n \to \infty$. Similar results are easily derived for left tails by considering $\{-X_i, i \geq 1\}$. The upper bound provided by Chernoff's bounding method (Chernoff, 1952) turns out to be the tight bound we are looking for. Let $\lambda > 0$, then

$$\mathbf{P}\{S_n \geq tn\} \;=\; \mathbf{P}\left\{e^{\lambda S_n} \geq e^{\lambda tn}\right\}.$$

It follows using Markov's inequality that

$$\mathbf{P}\{S_n \geq tn\} \;\leq\; e^{-\lambda tn} \cdot \mathbf{E}e^{\lambda S_n} \;=\; e^{-\lambda tn} \cdot \prod_{i=1}^{n} \mathbf{E}e^{\lambda X_i}, \tag{2.2}$$

since the variables $X_i$ are independent. The *cumulant generating function*, defined by $\Lambda(\lambda) = \log \mathbf{E}\left[e^{\lambda X}\right]$, plays an important role. Rewriting (2.2) using $\Lambda$, we obtain

$$\mathbf{P}\{S_n \geq tn\} \;\leq\; e^{-\lambda tn+\Lambda(\lambda)n}.$$

So optimizing our choice of $\lambda$, we see that

$$\mathbf{P}\{S_n \geq tn\} \leq \left(\inf_{\lambda>0} e^{-\lambda t+\Lambda(\lambda)}\right)^n \stackrel{\text{def}}{=} e^{-n\Lambda^\star(t)}, \tag{2.3}$$

where $\Lambda^\star(t) = \sup_\lambda\{\lambda t - \Lambda(\lambda)\}$ is the Fenchel–Legendre (convex) dual of $\Lambda$ (see, e.g., Rockafellar, 1970). The upper bound (2.3) is tight, as claimed by Cramér's theorem (Cramér, 1938).

**Theorem 2.1** (Cramér). *Assume that $\Lambda(\lambda) < \infty$ for some $\lambda > 0$. Let $t \geq \mathbf{E}X$. Then, as $n \to \infty$,*

$$\mathbf{P}\{S_n \geq tn\} = e^{-n\Lambda^\star(t)+o(n)}.$$

Cramér's version of the theorem was restricted to random variables on $\mathbb{R}$ having densities. The generalization is due to Chernoff (1952). See Petrov (1975) for more information. A complete proof can be found in Dembo and Zeitouni (1998).

The rate function $\Lambda^\star$ describing the tail probability in Theorem 2.2 is thus of great importance. Indeed, the behaviour of $\mathbf{P}\{S_n \geq tn\}$ relies directly on its properties. For instance, Theorem 2.2 would be useless if one cannot prove that $\Lambda^\star \geq 0$ and not identically zero. We review most useful properties of $\Lambda^\star$ in section 2.5. But first, we focus our attention to the generalizations of Cramér's theorem that we will need in the course of the proofs.

## 2.4 From Cramér to Gärtner–Ellis

This section is devoted to large deviations between the value of a sum of random *vectors* and its expected value. We are interested in the case of *extended* random vectors, that is, whose components may also take (only) one of the values $\infty$ or $-\infty$. We now focus on this slight generalization.

Let $\{X_i, 1 \leq i \leq n\}$ be a family of i.i.d. extended random vectors $X_i = (Z_i, E_i)$ distributed like $X = (Z, E)$. Assume $Z \in [-\infty, \infty)$ and $E \in [0, \infty]$. Set $p =$

$\mathbf{P}\{Z > -\infty, E < \infty\}$. For $\alpha$ and $\rho$ real numbers, we are interested in the tail probability

$$\mathbf{P}\left\{\sum_{i=1}^{n} Z_i > \alpha n, \sum_{i=1}^{n} E_i < \rho n\right\}, \tag{2.4}$$

whose magnitude is characterized in Cramér's theorem. We shall introduce the cumulant generating function $\Lambda$ of an *extended* random vector $X$. For $\lambda, \mu \in \mathbb{R}$, it is defined by

$$\Lambda(\lambda, \mu) = \log \mathbf{E}\left[\, e^{\lambda Z + \mu E} \;\middle|\; Z > -\infty, E < \infty \right] + \log p. \tag{2.5}$$

Observe that if $Z$ and $E$ are a.s. real, then $\Lambda(\lambda, \mu) = \mathbf{E}\left[e^{\lambda Z + \mu E}\right]$, which matches the usual definition. The tail probability in (2.4) is characterized using $\Lambda^\star$, the Fenchel–Legendre dual of $\Lambda$ (see Rockafellar, 1970): for $\alpha, \rho \in \mathbb{R}$, we define

$$\Lambda^\star(\alpha, \rho) = \sup_{\lambda,\mu}\{\lambda\alpha + \mu\rho - \Lambda(\lambda, \mu)\}.$$

**Theorem 2.2 (Cramér).** *Assume that $\{X_i, i \geq 1\}$ are i.i.d. random vectors distributed like $X$. Assume that $0 \in \mathcal{D}_\Lambda^\circ$. Let $I(\alpha, \rho) = \inf\{\Lambda^\star(x, y) : x > \alpha, y < \rho\}$. Then for any $\alpha, \rho \in \mathbb{R}$,*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left\{\sum_{i=1}^{n} Z_i > \alpha n, \sum_{i=1}^{n} E_i < \rho n\right\} = -I(\alpha, \rho).$$

*Moreover, the following explicit upper bound holds for all $n \geq 1$, and $\alpha, \rho \in \mathbb{R}$:*

$$\mathbf{P}\left\{\sum_{i=1}^{n} Z_i > \alpha n, \sum_{i=1}^{n} E_i < \rho n\right\} \leq e^{-nI(\alpha,\rho)}.$$

**Remarks.** (a) It is possible that $\Lambda^\star = \infty$ everywhere except in one point, and consequently $I$ may be infinite as well.

(b) Observe that the inequalities in Theorem 2.2 are strict. The result is false if one allows equality (see Groeneboom et al. (1979) or Dembo and Zeitouni (1992, Exercise 2.2.37) for a counterexample built by taking $(\alpha, \rho)$ on the boundary of $\mathcal{D}_\Lambda$). This technicality may be avoided if one enforces $(\alpha, \rho) \in \mathcal{D}_{\Lambda^\star}^\circ$ (see Lemma 2.2).

(c) The explicit upper bound is analogous to the Chernoff bound (Chernoff, 1952) and holds because the quadrant $(\alpha, \infty) \times (0, \rho)$ is a convex set (see Exercise 2.2.38, p. 42, Dembo and Zeitouni, 1998).

*Proof.* The quadrant $(\alpha, \infty) \times (-\infty, \rho)$ is a convex open set in $\mathbb{R}^2$. Hence Theorem 6.1.8 of Dembo and Zeitouni (1992) applies when $\mathbf{P}\{Z = -\infty \text{ or } E = \infty\} = 0$ (thus, $p = 1$). We now show the details in the extended case. Let $F_n = \{Z_i > -\infty, E_i < \infty, 1 \le i \le n\}$. It is the case that

$$\mathbf{P}\left\{\sum_{i=1}^n Z_i > \alpha n, \sum_{i=1}^n E_i < \rho n\right\} = \mathbf{P}\left\{\sum_{i=1}^n Z_i > \alpha n, \sum_{i=1}^n E_i < \rho n \;\middle|\; F_n\right\} \cdot p^n.$$

The classical form of Cramér's theorem applies to the first factor, and hence, writing $\Lambda_c = (\lambda, \mu) \mapsto \log \mathbf{E}\left[\, e^{\lambda Z + \mu E} \;\middle|\; Z > -\infty, E < \infty \right]$, the cumulant generating function of $(Z, E)$ conditioned on $\{Z > -\infty, E < \infty\}$, and $\Lambda_c^\star$ for its dual,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left\{\sum_{i=1}^n Z_i > \alpha n, \sum_{i=1}^n E_i < \rho n\right\} = -\inf\{\Lambda_c^\star(x, y) : x > \alpha, y < \rho\} + \log p.$$

However, $\Lambda = \Lambda_c + \log p$, and therefore $\Lambda^\star = \Lambda_c^\star - \log p$, which finishes the proof. □

The constraint that $\{X_i, i \ge 0\}$ be identically distributed may be relaxed, and we will need such an extension in Chapter 5. The case where the random variables are not identically distributed is treated by the Gärtner–Ellis theorem (Gärtner, 1977; Ellis, 1984) (actually the random variables need not be independent either). We will only use the upper bound. We shall first state the classical version of the Gärtner–Ellis theorem, and then extend it slightly to fit our needs.

**Theorem 2.3 (Gärtner–Ellis).** *Let $\{(Z_n, E_n), n \ge 1\}$ be random vectors taking values in $\mathbb{R} \times [0, \infty)$. Assume that for all $\lambda, \mu \in \mathbb{R}$,*

$$\Lambda(\lambda, \mu) = \lim_{n\to\infty} \frac{1}{n} \log \mathbf{E}\left[e^{\sum_{i=1}^n (\lambda Z_i + \mu E_i)}\right]$$

*exists as an element of $(-\infty, \infty]$. We assume that $\Lambda$ is the cumulant generating function of some random vector $X$. If $0 \in \mathcal{D}_\Lambda^\circ$, then*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left\{\sum_{i=1}^n Z_i > \alpha n, \sum_{i=1}^n E_i < \rho n\right\} \le -I(\alpha, \rho),$$

*where $I(\alpha, \rho) = \inf\{\Lambda^\star(x, y) : x > \alpha, y < \rho\}$, and $\Lambda^\star$ is the Fenchel-Legendre transform of $\Lambda$.*

A complete proof of Theorem 2.3 may be found in Dembo and Zeitouni (1998). Observe that Theorem 2.3 only requires *pointwise* convergence of the moment generating functions. We wish to extend the result slightly in order to handle extended random vectors, and obtain an explicit bound that does not involve limits. We shall also relax the assumption in order to require only asymptotic bounds on the generating function of the cumulants.

**Theorem 2.4 (Gärtner–Ellis).** *Let $\{(Z_n, E_n), n \geq 1\}$ be random vectors taking values in $[-\infty, \infty) \times [0, \infty]$. Let $F_n = \{Z_i > -\infty, E_i < \infty, 1 \leq i \leq n\}$. Let $A_m, m \geq 0$ be an arbitrary sequence of events. Assume that for all $(\lambda, \mu) \in \mathbb{R}^2$, and $\delta > 0$, there exists $M = M(\lambda, \mu, \delta)$ such that*

$$\sup_{n \geq 1} \frac{1}{n} \log \mathbf{E}\left[1[F_n, A_M] \cdot \exp\left(\sum_{i=1}^{n} \lambda Z_i + \mu E_i\right)\right] \leq \Lambda(\lambda, \mu) + \delta \leq \infty. \tag{2.6}$$

*We suppose that $\Lambda$ is the cumulant generating function of some vector $X$. Let $\Gamma$ be a closed set such that $\{-\infty\} \times [0, \infty] \cup [-\infty, +\infty) \times \{\infty\} \notin \Gamma$. Assume that $0 \in \mathcal{D}_\Lambda^o$. Then, for any $\gamma > 0$, there exists $M'$ such that*

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbf{P}\left\{\frac{1}{n}\sum_{i=1}^{n}(Z_i, E_i) \in \Gamma, A_{M'}\right\} \leq -\min\left\{1/\gamma, \inf_{(\alpha,\rho)\in\Gamma} \Lambda^\star(\alpha, \rho) - \gamma\right\},$$

*where $\Lambda^\star$ is the convex dual of $\Lambda$.*

In particular, when the set of interest is a quadrant, we have:

**Corollary 2.1.** *Let $\{(Z_n, E_n), n \geq 1\}$ be random vectors and $A_m$, $m \geq 1$ be events satisfying the conditions of Theorem 2.4. Then for every $(\alpha, \rho) \in \mathcal{D}_\Lambda^o$, and $\gamma > 0$, there exists $M'$ such that*

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbf{P}\left\{\sum_{i=1}^{n} Z_i > \alpha n, \sum_{i=1}^{n} E_i < \rho n, A_{M'}\right\} \leq -\min\{1/\gamma, I(\alpha, \rho) - \gamma\},$$

*where $I(\alpha, \rho) = \inf\{\Lambda^\star(x, y) : x > \alpha, y < \rho\}$, and $\Lambda^\star$ is the convex dual of $\Lambda$.*

*Proof of the Gärtner–Ellis theorem (Theorem 2.4).* The proof follows roughly the lines of that presented by Dembo and Zeitouni (1998). Let $\gamma > 0$. Observe first that, since

$\{-\infty\} \times [0, \infty] \cup [-\infty, +\infty) \times \{\infty\} \notin \Gamma$, we have, for all $M$,

$$\mathbf{P}\left\{\frac{1}{n}\sum_{i=1}^{n}(Z_i, E_i) \in \Gamma, A_M\right\} = \mathbf{P}\left\{\frac{1}{n}\sum_{i=1}^{n}(Z_i, E_i) \in \Gamma, A_M, F_n\right\}. \tag{2.7}$$

Reducing $\Gamma$ to a compact set. The first step consist in bounding (2.7) to a similar probability involving a compact set. Since $0 \in \mathcal{D}_\Lambda^\circ$, there exists $\lambda$ and $\mu > 0$ and $A > 0$ such that $\Lambda(\lambda, \mu) < A$. For any $r > 0$, we have

$$\mathbf{P}\left\{\sum_{i=1}^{n}\lambda Z_i + \mu E_i > rn, A_M, F_n\right\} \leq \mathbf{E}\left[\mathbf{1}[A_M, F_n] \cdot \exp\left(\sum_{i=1}^{n}\lambda Z_i + \mu E_i\right)\right] e^{-rn}.$$

Applying assumption (2.6) for this $\lambda$ and $\mu$, for all $M \geq M_1$ large enough, since $\Lambda(\lambda, \mu) \leq A$,

$$\mathbf{P}\left\{\sum_{i=1}^{n}\lambda Z_i + \mu E_i > rn, A_M, F_n\right\} \leq e^{(A+\delta-r)n}.$$

Therefore, for $r = A + \delta + 1/\gamma$, writing $C = \{(x, y) : \lambda x + \mu y \leq r\}$, and $\Gamma' = \Gamma \cap C$ we see that, for $M \geq M_1$,

$$\begin{aligned}\mathbf{P}\left\{\sum_{i=1}^{n}(Z_i, E_i) \in \Gamma, A_M, F_n\right\} &\leq \mathbf{P}\left\{\sum_{i=1}^{n}(Z_i, E_i) \in \Gamma', A_M, F_n\right\} \\ &\quad + \mathbf{P}\left\{\sum_{i=1}^{n}(Z_i, E_i) \in C^c, A_M, F_n\right\},\end{aligned}$$

and hence, for $M \geq M_1$,

$$\mathbf{P}\left\{\sum_{i=1}^{n}(Z_i, E_i) \in \Gamma, A_M, F_n\right\} = \mathbf{P}\left\{\sum_{i=1}^{n}(Z_i, E_i) \in \Gamma', A_M, F_n\right\} + e^{-n/\gamma}. \tag{2.8}$$

Covering $\Gamma'$ with small sets. We now proceed by covering $\Gamma'$ with a finite set of balls. For any $\omega = (x_\omega, y_\omega) \in \mathbb{R}^2$, there exists $(\lambda_\omega, \mu_\omega)$ such that

$$\lambda_\omega x_\omega + \mu_\omega y_\omega - \Lambda(\lambda_\omega, \mu_\omega) > \min\left\{\frac{1}{\gamma} + \frac{2\gamma}{3}, \Lambda^*(x_\omega, y_\omega) - \frac{\gamma}{3}\right\}.$$

For all $\omega \in \mathbb{R}^2$, there exists an open ball $\mathcal{B}_\omega$ such that for all $(x, y) \in \mathcal{B}_\omega$, $|\lambda_\omega(x - x_\omega) + \mu_\omega(y - y_\omega)| \leq \gamma/3$. Hence we have

$$\inf_{(x,y)\in\mathcal{B}_\omega}\{\lambda_\omega x + \mu_\omega y\} \geq \Lambda(\lambda_\omega, \mu_\omega) - \frac{\gamma}{3} + \min\left\{\frac{1}{\gamma} + \frac{2\gamma}{3}, \Lambda^*(x_\omega, y_\omega) - \frac{\gamma}{3}\right\}. \tag{2.9}$$

The set $\{\mathcal{B}_\omega, \omega \in \mathbb{R}^2\}$ covers $\mathbb{R}^2$ but is uncountable. However $\Gamma'$ is contained in a compact set, and it can be covered by $\{\mathcal{B}_\omega, \omega \in \mathcal{C}\}$, where $\mathcal{C}$ is finite. Thus, by the union bound,

$$\begin{aligned}\mathbf{P}\left\{\sum_{i=1}^n (Z_i, E_i) \in n\Gamma', A_M, F_n\right\} &\le \mathbf{P}\left\{\frac{1}{n}\sum_{i=1}^n (Z_i, E_i) \in \bigcup_{\omega\in\mathcal{C}} \mathcal{B}_\omega, A_M, F_n\right\} \\ &\le \sum_{\omega\in\mathcal{C}} \mathbf{P}\left\{\frac{1}{n}\sum_{i=1}^n (Z_i, E_i) \in \mathcal{B}_\omega, A_M, F_n\right\}.\end{aligned}$$

Consider one term in the sum above. Note that

$$\sum_{i=1}^n (Z_i, E_i) \in \mathcal{B}_\omega \quad\Rightarrow\quad \sum_{i=1}^n \lambda_\omega Z_i + \mu_\omega E_i \ge \inf_{(x,y)\in\mathcal{B}_\omega} \{\lambda_\omega x + \mu_\omega y\}.$$

Then, using assumption (2.6) with $\lambda_\omega$ and $\mu_\omega$ and $\delta = \gamma/3$, there exists $M_2 = M_2(\omega)$ such that for all $M \ge M_2$,

$$\mathbf{P}\left\{\sum_{i=1}^n (Z_i, E_i) \in \mathcal{B}_\omega, A_M\right\} \le \exp\left(n\Lambda(\lambda_\omega, \mu_\omega) + n\frac{\gamma}{3} - n \inf_{(x,y)\in\mathcal{B}_\omega} \{\lambda_\omega x + \mu_\omega y\}\right).$$

Then, recalling the bound (2.9), we obtain for all $M \ge M_2(\omega)$,

$$\mathbf{P}\left\{\sum_{i=1}^n (Z_i, E_i) \in \mathcal{B}_\omega, A_M\right\} \le \exp\left(n\frac{2\gamma}{3} - n \cdot \min\left\{\frac{1}{\gamma} + \frac{2\gamma}{3}, \Lambda^\star(x_\omega, y_\omega) - \frac{\gamma}{3}\right\}\right).$$

Finally, plugging the bound above in (2.8), and observing that $\inf\{\Lambda^\star(x,y) : (x,y) \in \Gamma\} = I(\alpha, \rho)$, for all $M \ge \max\{M_1, M_2(\omega) : \omega \in \mathcal{C}\}$

$$\mathbf{P}\left\{\frac{1}{n}\sum_{i=1}^n (Z_i, E_i) \in \Gamma, A_M\right\} \le (1 + |\mathcal{C}|) \cdot \exp\left(-n \cdot \min\left\{\frac{1}{\gamma}, I(\alpha, \rho) - \gamma\right\}\right).$$

Taking logarithms completes the proof. □

## 2.5 About $\Lambda$, $\Lambda^\star$ and $I$

The functions $\Lambda$, $\Lambda^\star$ and $I$ are well understood (Dembo and Zeitouni, 1992). They will be the corner stone of the characterization of first order asymptotic properties of the height of random trees. This is why we collect here their main properties.

Consider a mapping $f : \mathbb{R}^2 \to (-\infty, \infty]$. Recall that $\mathcal{D}_f$ is its domain: $\mathcal{D}_f = \{(\alpha, \rho) : f(\alpha, \rho) < \infty\}$, and $\mathcal{D}_f^o$ is the interior of $\mathcal{D}_f$.

The mapping $f$ is said to be *convex* if, for all $x_1, x_2 \in \mathbb{R}^2$, and $\theta \in [0, 1]$, we have $f(\theta x_1 + (1-\theta)x_2) \le \theta f(x_1) + (1-\theta) f(x_2)$, where it is understood that if the left-hand side is $\infty$, then either $f(x_1) = \infty$ or $f(x_2) = \infty$. If the *level sets* $\Psi_f(\ell) = \{x : f(x) \le \ell\}$ are closed for all $\ell \in \mathbb{R}$, we say that $f$ is *lower semicontinuous*, and call $f$ a *rate function*. The mapping $f$ is said to be a *good rate function* if its level sets are compact.

THE FUNCTION $\Lambda(\cdot,\cdot)$. The cumulant generating function is the link between the random variables and the rate functions, and its properties imply those of $\Lambda^\star$ and $I$.

**Lemma 2.1.** *The function* $\Lambda(\cdot,\cdot)$
*(a) takes values in* $(-\infty, \infty]$ *if* $p = \mathbf{P}\{Z > -\infty, E < \infty\} > 0$*;*
*(b) is convex on* $\mathbb{R}^2$*, and continuous in* $\mathcal{D}_\Lambda^o$.

*Proof.* (a) By definition, $\forall \lambda, \mu \in \mathbb{R}$, we have

$$\Lambda(\lambda, \mu) = \log p + \log \mathbf{E}\left[\, e^{\lambda Z + \mu E} \;\middle|\; Z > -\infty, E < \infty \right].$$

Both $Z$ and $E$ are real on $\{Z > \infty, E < \infty\}$, and hence $\mathbf{E}\left[e^{\lambda Z + \mu E}\right] > 0$. Since $p > 0$, this yields $\Lambda(\lambda, \mu) > -\infty$.

(b) The convexity follows from Hölder's inequality. The continuity in $\mathcal{D}_\Lambda^o$ is a straighforward consequence of the convexity. For details see Dembo and Zeitouni (1992). $\square$

THE FUNCTION $\Lambda^\star(\cdot,\cdot)$. The level sets of $\Lambda^\star$ are of particular interest, and we write $\Psi = \Psi_{\Lambda^\star}$. Indeed, as we will see in Chapters 4 to 6, the heights will be characterized using optimizations of some objective functions on the level sets of $\Lambda^\star$.

**Lemma 2.2.** *The function* $\Lambda^\star(\cdot,\cdot)$ *is*
*(a) convex on* $\mathbb{R}^2$ *;*
*(b) continuous on* $\mathcal{D}_{\Lambda^\star}^o$*;*

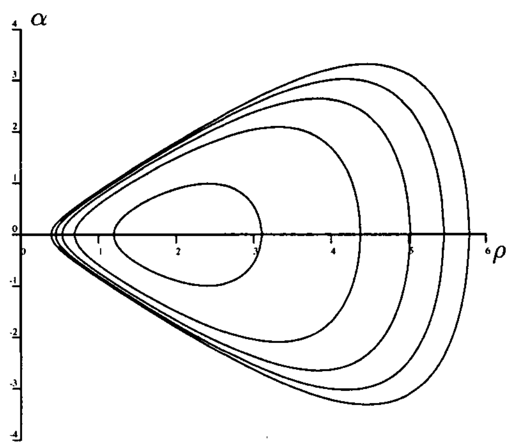


*Figure 2.1: An increasing family of level sets $\Psi(\ell)$ for the function $\Lambda^\star$ corresponding to the study of k-d trees (Section 5.7.12)*

*(c) a good rate function if* $0 \in \mathcal{D}_\Lambda^o$.

*Proof.* (a) The convexity of $\Lambda^\star$ is a direct consequence of its definition: for $t_1, t_2, \lambda \in \mathbb{R}^2$ and $\theta \in [0,1]$, using $\cdot$ to denote the standard inner product,

$$\begin{aligned}
\Lambda^\star(\theta t_1 + (1-\theta)t_2) &= \sup_{\lambda\in\mathbb{R}^2}\{\lambda\cdot(\theta t_1 + (1-\theta)t_2) - \Lambda(\lambda)\} \\
&\le \sup_{\lambda\in\mathbb{R}^2}\{\theta\lambda\cdot t_1 - \theta\Lambda(\lambda)\} + \sup_{\lambda\in\mathbb{R}^2}\{(1-\theta)\lambda\cdot t_2 - (1-\theta)\Lambda(\lambda)\} \\
&= \theta\Lambda^\star(t_1) + (1-\theta)\Lambda^\star(t_2).
\end{aligned}$$

(b) Since $\Lambda^\star$ is convex, it is continuous on $\mathcal{D}_{\Lambda^\star}^o$.
(c) Let $\ell \ge 0$. For $r \ge 0$, let $\mathscr{C}_r = \{(x,y) \in \mathbb{R}^2 : \sqrt{x^2+y^2} \ge r\}$. Since $0 \in \mathcal{D}_\Lambda^o$, there exists a ball $\mathcal{B}$, centered at the origin with radius $\delta > 0$, and $A < \infty$ such that for all $(\lambda,\mu) \in \mathcal{B}$, $\Lambda(\lambda,\mu) \le A$. For any $r \ge 0$ and $(\alpha,\rho) \in \mathscr{C}_r$,

$$\Lambda^\star(\alpha,\rho) = \sup_{\lambda,\mu}\{\lambda\alpha + \mu\rho - \Lambda(\lambda,\mu)\} \ge \sup_{(\lambda,\mu)\in\mathcal{B}}\{\lambda\alpha + \mu\rho - \Lambda(\lambda,\mu)\} \ge \delta\cdot r - A.$$

As a consequence, for $R$ large enough, $\mathscr{C}_R$ does not intersect $\Psi(\ell)$, proving that $\Psi(\ell)$ is bounded.

We now show that $\Psi(\ell)$ is closed ($\Lambda^\star$ is lower semicontinuous). It suffices to prove that $\Psi(\ell)$ contains all its accumulation points: for any $(\alpha,\rho) \in \mathbb{R}^2$ such that there exists $(\alpha_n,\rho_n) \in \Psi(\ell)$ with $(\alpha_n,\rho_n) \to (\alpha,\rho)$, we should have $(\alpha,\rho) \in \Psi(\ell)$. For any $\lambda,\mu \in \mathbb{R}$,

$$\liminf_{n\to\infty}\Lambda^\star(\alpha_n,\rho_n) \ge \liminf_{n\to\infty}\{\lambda\alpha_n + \mu\rho_n - \Lambda(\lambda,\mu)\} = \lambda\alpha + \mu\rho - \Lambda(\lambda,\mu).$$

As a result,

$$\liminf_{n\to\infty} \Lambda^\star(\alpha_n, \rho_n) \geq \sup_{\lambda,\mu}\{\lambda\alpha + \mu\rho - \Lambda(\lambda,\mu)\} = \Lambda^\star(\alpha,\rho).$$

Hence, $\Lambda^\star(\alpha,\rho) \leq \ell$ and $(\alpha,\rho) \in \Psi(\ell)$, which proves that $\Psi(\ell)$ is closed. □

THE FUNCTION $I(\cdot,\cdot)$. The function that appears in the Cramér and Gärtner–Ellis theorems is $I(\cdot,\cdot)$. This is why information about tail probabilities rely on properties of $I$.

**Lemma 2.3.** *For $\alpha, \rho \in \mathbb{R}^2$, let $I(\alpha,\rho) \overset{def}{=} \inf\{\Lambda^\star(x,y) : x > \alpha, y < \rho\}$. Then*
*(a) $(\alpha,\rho) \mapsto I(\alpha,\rho)$ is non-decreasing in $\alpha$, and non-increasing in $\rho$;*
*(b) for $(\alpha,\rho) \in \mathcal{D}^\circ_{\Lambda^\star}$, $I(\alpha,\rho) = \inf\{\Lambda^\star(x,y) : x \geq \alpha, y \leq \rho\}$.*

*Proof.* (a) This is clear from the definition as an infimum.

(b) Clearly $\inf\{\Lambda^\star(x,y) : x \geq \alpha, y \leq \rho\} \leq I(\alpha,\rho)$. So we prove that $I(\alpha,\rho) \leq \inf\{\Lambda^\star(x,y) : x \geq \alpha, y \leq \rho\}$. Consider a sequence $(x_n, y_n) \in \mathcal{D}_{\Lambda^\star}$ such that

$$\lim_{n\to\infty} \Lambda^\star(x_n,y_n) = \inf\{\Lambda^\star(x,y) : x \geq \alpha, y \leq \rho\}.$$

We build an auxiliary sequence $(x'_n, y'_n)$, $n \geq 1$. Let $k \geq 1$. If $x_k \neq \alpha$ and $y_k \neq \rho$, then $(x'_k, y'_k) = (x_k, y_k)$. Assume now that $x_k = \alpha$ or $y_k = \rho$. Then we construct a new point $(x'_k, y'_k)$ such that $\Lambda^\star(x'_k, y'_k) < \Lambda^\star(x_k, y_k) + 1/k$ where $x'_k > \alpha$ and $y'_k < \rho$. This construction is done in the following way. Assume at first that for small enough $\epsilon > 0$ there exists a ball $\mathcal{B}_\epsilon$ centered at $(x_k, y_k)$ with radius $\epsilon$ contained within $\mathcal{D}^\circ_{\Lambda^\star}$. In this case, by the continuity of $\Lambda^\star$, we find a point $(x'_k, y'_k) \in \mathcal{B}_\epsilon$ with $x_k > \alpha, y_k < \rho$ such that $\Lambda^\star(x'_k, y'_k) < \Lambda^\star(x_k, y_k) + 1/k$. In the second case, no such ball exists for any $\epsilon$, which means in particular $(x_k, y_k)$ lies on the boundary of $\mathcal{D}_{\Lambda^\star}$. Consider the region $R_\epsilon = \mathcal{B}_\epsilon \cap \mathcal{D}^\circ_{\Lambda^\star} \cap \{(\alpha,\infty) \times (-\infty,\rho)\}$. Since $(\alpha,\rho) \in \mathcal{D}^\circ_{\Lambda^\star}$, an open convex set, this region is non-empty. Let $\beta = \inf_{\epsilon\downarrow 0} \sup\{\Lambda^\star(x,y) : (x,y) \in R_\epsilon \setminus (x_k,y_k)\}$. Assume for a contradiction that $\beta > \Lambda^\star(x_k,y_k)$. Then there exist $(x,y)$ such that the line joining $(x,y)$ to $(x_k,y_k)$ lies below $\Lambda^\star$, contradicting the convexity of $\Lambda^\star$. Hence $\beta \geq \Lambda^\star(x_k,y_k)$ and, for $\epsilon$ small enough, there exist $(x'_k, y'_k)$ in $R_\epsilon$ such that $\Lambda^\star(x'_k, y'_k) \leq \Lambda^\star(x_k, y_k)$.

Therefore, using the auxiliary sequence, we see that $\inf\{\Lambda^\star(x,y) : x > \alpha, y < \rho\} \leq \lim_{n\to\infty} \Lambda^\star(x'_n, y'_n) = \inf\{\Lambda^\star(x,y) : x \geq \alpha, y \leq \rho\}$. This finishes the proof. □

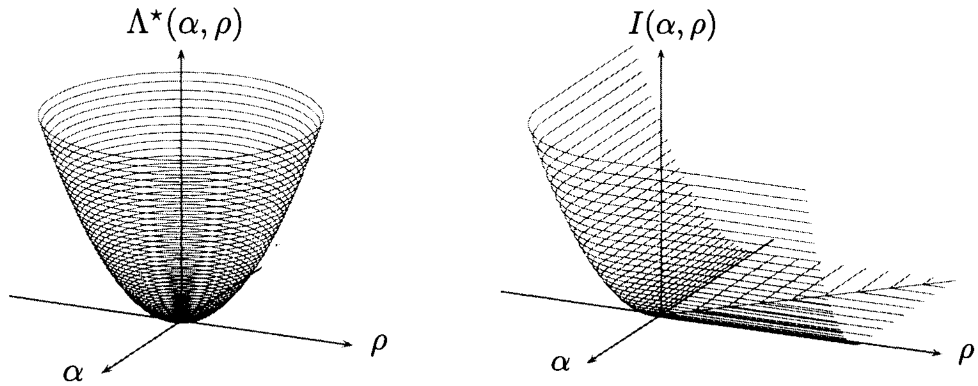


***Figure 2.2:*** *The functions $\Lambda^\star(\alpha,\rho)$ (left), $I(\alpha,\rho)$ (right) for the example of the Gaussian random variables.*

**Example: Gaussian random variables.** We now just work out an easy example to emphasize the differences between $\Lambda$, $\Lambda^\star$ and $I$. Assume that $(Z, E)$ is distributed as $(N_1, N_2)$, where $N_1$ and $N_2$ are independent standard Gaussian random variables $\mathcal{N}(0,1)$. Then, for all $\lambda, \mu \in \mathbb{R}$,

$$\Lambda(\lambda,\mu) = \log \mathbf{E}\left[e^{\lambda Z}\right] + \log \mathbf{E}\left[e^{\mu E}\right].$$

Furthermore,

$$\mathbf{E}\left[e^{\lambda Z}\right] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\lambda z - z^2/2} dz = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(z-\lambda)^2/2}\cdot e^{\lambda^2/2} dx = e^{\lambda^2/2}.$$

A similar statement holds clearly about $E$. It follows that

$$\Lambda(\lambda,\mu) = \frac{\lambda^2}{2} + \frac{\mu^2}{2}.$$

Then, the optimum in the definition of $\Lambda^\star(\alpha,\rho)$ is obtained for $\lambda = \alpha$ and $\mu = \rho$, and we have

$$\Lambda^\star(\alpha,\rho) = \frac{\alpha^2}{2} + \frac{\rho^2}{2}.$$

The definition of $I(\alpha, \rho)$ then depends on where $(\alpha, \rho)$ lies with respect to $(0, 0)$:

$$I(\alpha,\rho) = \begin{cases} \Lambda^\star(\alpha,\rho) & \alpha > 0, \rho < 0, \\ \alpha^2/2 & \alpha > 0, \rho > 0, \\ \rho^2/2 & \alpha < 0, \rho < 0, \\ 0 & \alpha < 0, \rho > 0. \end{cases}$$

Figure 2.2 shows the functions $\Lambda^\star(\alpha, \rho)$ and $I(\alpha, \rho)$ for this example.

*Chapter 3*

# Branching Processes

*In this chapter, we introduce the theory of branching processes. The Galton–Watson process is the simplest of all branching processes. We review branching random walks since most of the work in this thesis is tightly connected to a generalized version of the first-birth problem. For further information, see the textbooks of Harris (1963), Athreya and Ney (1972), Jagers (1975). Devroye (1998a) surveys their applications in analysis of algorithms.*

*Trees are sanctuaries. Whoever knows how to talk to them, whoever knows how to listen to them, can learn the truth.*
– Hermann Hesse

## Contents

## 3.1 The Galton–Watson process

### 3.1.1 Definition and main results

In the late nineteenth century, F. Galton became interested in the decay of family names in England (Kendall, 1966). He was mainly concerned with "men of the note". He formalized his problem mathematically and communicated it in the following way Galton (1873):

> *A large nation, of whom we will only concern ourselves with adult males, $N$ in number, and who each bear separate surnames, colonise a district. Their law of population is such that, in each generation, $a_0$ per cent of the adult males have no male children who reach adult life; $a_1$ have one such male child; $a_2$ have two; and so on up to $a_5$ who have five.*
>
> *Find (1) what proportion of the surnames will have become extinct after $r$ generations; and (2) how many instances there will be of the same surname being held by $m$ persons.*

He was not pleased with the only solution he was proposed and urged Reverend H.W. Watson, whom he was corresponding with, to take up the matter. Watson made use of generating functions and functional iterations to tackle the problem. The following approach is essentially his. If we write $p_k$ for Galton's $a_k$, and remove the restriction that $k \leq 5$, we can define the *probability generating function* $f$ associated with the distribution $\{p_i, i \geq 0\}$ by

$$f(s) = \sum_{k=0}^{\infty} p_k s^k,$$

for $s \in [0,1]$. Also, if we introduce the $n$-fold convolution of $f$ with itself, $f_1 = f$, $f_{n+1} = f \circ f_n = f_n \circ f$, then the coefficients of the power series for $f_n$ are the terms of the probability distribution for the number of males in the $n$-th generation. Galton observed that the *probability of extinction* by the $n$-th generation, $q_n$, satisfies the following equations $q_1 = p_0$, $q_{n+1} = f(q_n)$ and if $q_n \to q$, then $f(q) = q$. This last equation accepting always 1 as a root, Galton inferred incorrectly that the male line's extinction was inevitable (see Galton and Watson, 1874).

The problem reappeared apparently independently in the late 1920s and Haldane (1927) and Steffensen (1930) finally stated correctly the Criticality Theorem. In today's words, we can define a Galton–Watson (GW) process in the following way. There is a single ancestor who gives birth to a random number $Z$ of new individuals according to a specified probability distribution. Further, any individual of the process reproduces similarly and independently. For $k \geq 0$, write $p_k = \mathbf{P}\{Z = k\}$. Then the *reproduction generating* function used by Watson is simply defined by

$$f(s) = \mathbf{E}\left[s^Z\right] = \sum_{k \geq 0} p_k s^k,$$

for $s \in [0,1]$. This function is convex (as a sum of convex functions), strictly convex if $p_1 \neq 1$, and increases from 0 to 1 for $s \in [0,1]$. This function concentrates all the information of the distribution of $Z$. In particular, the expected number of children $\mathbf{E}Z$ is

$$m \stackrel{\text{def}}{=} \mathbf{E}Z = \sum_{k \geq 0} k p_k = f'(1).$$

If we let $Z_n$ denote the number of individuals present in generation $n$, then, $f_n(s) = \mathbf{E}s^{Z_n}$. The extinction probability is defined by

$$q = \mathbf{P}\{\exists n : Z_n = 0\}.$$

**Theorem 3.1.** *Consider a Galton–Watson process with offspring distribution $Z$. Then $q$ is the smallest fixed point of the reproduction generating function in $[0,1]$. In particular, $q < 1$ if and only if $m = \mathbf{E}Z > 1$ or $\mathbf{E}Z = 1$ and $p_1 = 1$.*

This leads to the classification of Galton–Watson processes in three groups depending on the value of $m$: a process is called *subcritical*, *critical*, or *supercritical* if $m < 1$, $m = 1$ or $m > 1$ respectively. It is also very useful to observe that if the process does not become extinct, then size of the generation grows to infinity. In other words, with probability 1, $Z_n$ does not oscillate. This is one of the key results used in some of our lower bounds.

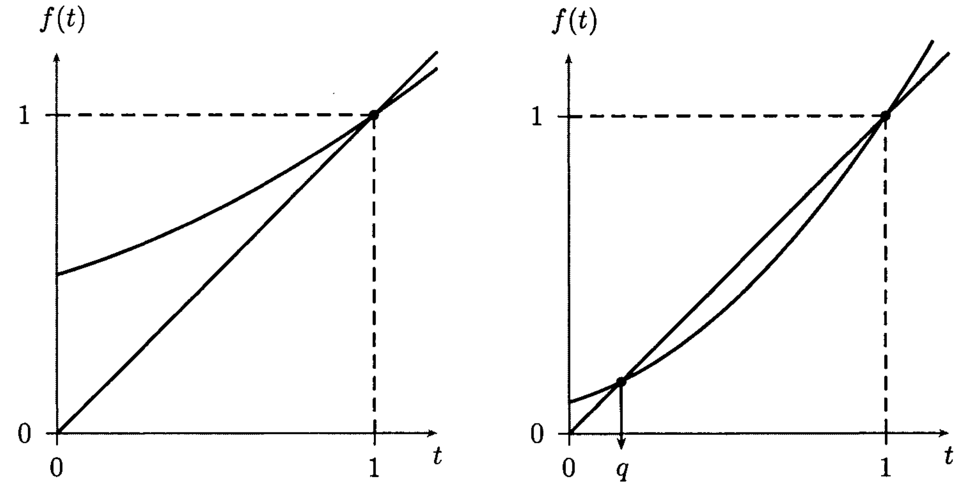


***Figure 3.1:*** *The generating functions for a subcritical (left) and supercritical (right) GW processes are shown. In the subcritical case, 1 is the only root of $f(t) = t$ in $[0,1]$. In the supercritical case, there is $q \in [0,1)$ such that $f(q) = q$.*

**Theorem 3.2.** *Let $Z_n$ be the number of individuals in the $n$-th generation of a Galton–Watson process with offspring distribution $Z$. Assume that $p_1 = \mathbf{P}\{Z = 1\} < 1$. Then $\lim_{n\to\infty} Z_n \in \{0, \infty\}$ almost surely.*

By a simple conditioning argument, one sees that $\mathbf{E}Z_n = m^n$. Indeed, clearly, $\mathbf{E}Z_0 = 1$ and proceeding by induction on $n$,

$$\mathbf{E}Z_n = \mathbf{E}\mathbf{E}\left[\, Z_n \mid Z_{n-1} \,\right] = \mathbf{E}\left[mZ_{n-1}\right] = m^n.$$

It actually turns out that $Z_n$ behaves roughly like $m^n$. Doob's limit law (see Harris, 1963) characterizes more precisely the behavior of $Z_n$.

**Theorem 3.3.** *Let $m$ be finite. Then, $W_n = Z_n/m^n$ forms a martingale with $\mathbf{E}W_n = 1$ and $W_n \to W$ almost surely, as $n \to \infty$, where $W$ is a nonnegative random variable.*

The distribution of $W$ is not known. However, one can obtain fairly precise information on $W$, and the process behaves *exactly* as one expects ($\mathbf{E}W = 1$, $\mathbf{P}\{W = 0\} = q$) if and only if the $x \log x$ moment of $Z$ is finite, as stated by the following theorem, due to Kesten and Stigum (1966), which pins down the asymptotic properties of supercritical Galton–Watson processes.

**Theorem 3.4** (Kesten and Stigum, 1966). *Let $Z_n$ the number of individuals in the $n$-th generation of a supercritical Galton–Watson process with progeny distribution $Z$. The following statements are equivalent:*

*(a)* $\lim_{n\to\infty} \mathbf{E}|W_n - W| = 0$ ;

*(b)* $\mathbf{E}[Z\log(1+Z)] < \infty$ ;

*(c)* $\mathbf{E}W = 1$ ;

*(d)* $\mathbf{P}\{W = 0\} = q$.

### 3.1.2 Bounding the extinction probability

Apart from the standard results we have just presented, we will need to bound the extinction probabilities of some Galton–Watson processes to boost some of our lower bounds.

**Theorem 3.5.** *Let $d \geq 1$ be a fixed integer. Consider a sequence of Galton–Watson processes with progeny distributions $Z^{(x)}$ on $\{0, 1, \ldots, d\}$, $\mathbf{E}Z^{(x)} = \mu_x$, and extinction probabilities $q_x$, $x \in \mathbb{R}$. Assume that there exists $x_0$ and $\delta > 0$ such that $\inf_{x\geq x_0} \mathbf{E}Z^{(x)} \geq 1+\delta$. If $\mathbf{P}\left\{Z^{(x)} = 0\right\} \to 0$, as $x \to \infty$, then $q_x \to 0$.*

**Remarks.** Before we proceed with the proof, observe that the result is best possible. Indeed, if either the support of $Z^{(x)}$ is unbounded or $\mathbf{E}Z^{(x)}$ is not uniformly bounded away from 1, one can construct distributions $Z^{(x)}$ for which the result does not hold. Write $p_i = p_i(x) = \mathbf{P}\left\{Z^{(x)} = i\right\}$, and let $f^{(x)}$ be the associated probability generating function: $\forall s$, $f^{(x)}(s) = \sum_{i\geq 0} p_i s^i$.

(a) We first build $Z^{(x)}$ such that $\mathbf{E}Z^{(x)} > 1$ for all $x \geq 2$, and yet $q_x \not\to 0$. Let

$$p_0 = \frac{1}{2x}, \qquad p_1 = 1 - \frac{2}{x}, \qquad p_2 = \frac{3}{2x}.$$

Then $\mathbf{E}Z^{(x)} = 1 + 1/x > 1$, for all $x \geq 2$. But $q_x$ is the smallest solution of $f^{(x)}(q_x) = q_x$. Thus for all $x \geq 2$, $q_x = 1/3 \not\to 0$.

(b) For the case of distributions with unbounded support, we consider, for $x \geq 2$ taking integer values only,

$$p_0 = \frac{1}{x}, \qquad p_1 = 1 - \frac{2}{x}, \qquad p_x = \frac{1}{x}.$$

Then for all $x \geq 4$, $\mathbf{E}Z^{(x)} = 2 - 2/x \geq 3/2$. However, $q_x$ is the smallest solution of $1 - 2q_x + q_x^n = 0$, and hence $q_x \to 1/2$ as $x \to \infty$.

The proof of Theorem 3.5 is based on the following Lemma providing an explicit bound on the extinction probability.

**Lemma 3.1.** *Let $d \geq 1$ be a fixed integer. Consider a Galton–Watson process with progeny distribution $Z$ on $\{0, 1, \ldots, d\}$ and extinction probability $q$. Let $\mu = \mathbf{E}Z$ and $p_i = \mathbf{P}\{Z = i\}$, $0 \leq i \leq d$. Assume that $p_1 < 1$. Then,*

$$q \leq \begin{cases} \dfrac{2p_0}{p_0 + \frac{\mu-1}{2}} & \text{if } \dfrac{2p_0}{p_0 + \frac{\mu-1}{2}} < 1 - \mu^{-\frac{1}{d-1}}, \\ 1 & \text{otherwise.} \end{cases}$$

*Proof.* The proof is based on an analysis of the probability generating function. We know that $q$ satisfies $f(q) = q$. Observe that

$$\mu = p_1 + \sum_{i \geq 2} i p_i \geq p_1 + 2 \sum_{i \geq 2} p_i.$$

If $\mu \leq 1$ then $q = 1$ by Theorem 3.1 and the result clearly holds. We assume now that $\mu > 1$ and $p_0 + p_1 < 1$. Define the auxiliary generating function

$$g(s) = \frac{f(s) - p_0 - p_1 s}{1 - p_0 - p_1},$$

and note that $g(s) \leq s^2$ for all $s \in [0, 1]$, and $g(0) = 0$. Now,

$$g(q) = \frac{f(q) - p_0 - p_1 q}{1 - p_0 - p_1} = \frac{q(1 - p_1) - p_0}{1 - p_0 - p_1} \leq q^2.$$

We rewrite the above equation in order to obtain a bound that is more useful

$$q \leq q^2 \cdot \frac{1 - p_1 - p_0}{1 - p_1} + \frac{p_0}{1 - p_1} \leq q^2 + \frac{p_0}{p_0 + \sum_{i>1} p_i} \leq q^2 + \frac{p_0}{p_0 + \frac{\mu - p_1}{2}}.$$

Finally, we see that

$$q \le q^2 + \frac{p_0}{p_0 + \frac{\mu-1}{2}} \stackrel{\text{def}}{=} q^2 + \alpha \quad \text{where} \quad \alpha = \frac{p_0}{p_0 + \frac{\mu-1}{2}}.$$

If $4\alpha \ge 1$, then clearly $q \le 1 \le 4\alpha$. Otherwise, $4\alpha < 1$ and this implies either

$$q \le \frac{1-\sqrt{1-4\alpha}}{2} \quad \text{or} \quad q \ge \frac{1+\sqrt{1-4\alpha}}{2}.$$

For all $x \in [0,1]$, we have $\sqrt{1-x} \ge 1-x$ and thus, we can conclude that, when $4\alpha < 1$, either

$$q \le 2\alpha \qquad \text{or} \qquad q \ge 1-2\alpha. \tag{3.1}$$

We now assume that $q \ne 0$, otherwise, the result trivially holds. Note that, in this case, since $q = f(q)$, $q \le \sum_{i=1}^{d} ip_i q^{i-1}$. By monotinicity, for the solution $r$ of $1 = \sum_{i=1}^{d} ip_i r^{i-1}$, we have $q \le r$. Observe also that $\sum_{i=1}^{d} ip_i = \mu$, which we have assumed greater than 1. As a consequence, $r \le 1$ and

$$1 \;=\; \sum_{i=1}^{d} ip_i r^{i-1} \;\ge\; r^{d-1}\sum_{i=1}^{d} ip_i \;=\; \mu r^{d-1},$$

so $q \le r \le \mu^{-\frac{1}{d-1}}$. Recalling (3.1), if $1-2\alpha > \mu^{-\frac{1}{d-1}}$, then we must have $q \le 2\alpha$. This proves the lemma. □

We are now ready to prove Theorem 3.5 which is, in fact, an easy corollary of Lemma 3.1.

*Proof of Theorem 3.5.* If $\mathbf{P}\left\{Z^{(x)} = 0\right\} = 0$, the result is clear. Assuming $\mathbf{P}\left\{Z^{(x)} = 0\right\} > 0$, recall Lemma 3.1. We have, for $x \ge x_0$,

$$1 - \mu_x^{-\frac{1}{d-1)}} \;\ge\; 1 - (1+\delta)^{-\frac{1}{d-1}} \stackrel{\text{def}}{=} \xi > 0.$$

As a consequence, since $p_0 = \mathbf{P}\left\{Z^{(x)} = 0\right\} \to 0$, for $x$ large enough,

$$\frac{2p_0}{p_2 + \frac{\mu_x - 1}{2}} \;=\; \frac{2p_0}{p_0 + \delta/2} \;\le\; \frac{4p_0}{\delta} \;<\; \xi.$$

Therefore, for $x \to \infty$, we have $q_x = O(p_0) = o(1)$, which completes the proof. □

### 3.1.3 Beyond Galton–Watson processes

In the course of the proofs, we will need the following technical lemma. One should see it as a tool to deal with branching processes for which the progeny distribution may depend on the node. It asserts that if there is a deterministic lower bound for the reproduction distribution function, then one can find a subprocess that is a proper Galton–Watson process, that is, for which every node has the same progeny distribution.

**Lemma 3.2.** *Let $N$ be a random positive integer, and given $N = n$, let $Z$ be a random variable distributed like $Z^{(n)}$, where,*

$$\inf_n \mathbf{P}\left\{Z^{(n)} \geq k\right\} \geq t_k$$

*and $t_k \downarrow 0$ as $k \to \infty$. Then one can find a random variable $Y$ such that $Y \leq Z$ and $\mathbf{P}\{Y \geq k\} = t_k$ for all $k$.*

*Proof.* Let $W$ be a random variable with tail distribution $t_k$: $\mathbf{P}\{W \geq k\} = t_k$. Let $F_n$ be the distribution function of $Z^{(n)}$ and $G$ be the distribution function of $W$. Let $U$ be uniformly distributed on $[0,1]$, then we couple $W$ and $\{Z^{(n)}, n \geq 0\}$ using the inverse transform technique (Billingsley, 1995; Grimmett and Stirzaker, 2001)

$$Z^{(n)} = F_n^{-1}(U), \quad \text{and} \quad Y = G^{-1}(U).$$

It is easy to see that with probability one, $W \leq Z^{(n)}$ for all $n$ and thus $W \leq Z$. $\square$

## 3.2 The first-birth problem

The first-birth problem is at the heart of the probabilistic branching processes techniques used in problems about the heights of trees. We make a brief overview of the main results, and present some theorems that will be useful later.

### 3.2.1 Discrete time branching random walks

The branching random walk (BRW) is one of the branching processes that models the growth of random trees. As its name indicates, it has a tree structure, with branching at each generation. Every individual bears a position on the real line $\mathbb{R}$, and every path in the tree is a random walk on $\mathbb{R}$.

More formally, there is an initial individual called the *ancestor*. The ancestor gives birth to an offspring with positions given by a real point process $Z$. Let $\{z_r^{(1)}, r \geq 1\}$ be the positions of the individuals in the first generation. For $n \geq 1$, assume we know the positions of all individuals up to the $n$-th generation. Then each one of these individuals reproduces in the same way as the ancestor, and independently of one another and of the past. More precisely, an individual with position $x$ gives birth to new individuals in the next generation, and their positions are distributed like $\{x + z_r^{(1)}, r \geq 1\}$. We write $\{z_r^{(n)}, r \geq 1\}$ for the positions of the people in the $n$-th generation. If $Z$ is concentrated on $[0, \infty)$ one can easily interpret the positions as time and consider the individuals that were born before some time $t$. One can keep this definition even if the walks down the ancestor are not monotonic, i.e., if $Z$ is not compelled to be nonnegative. Let $U$ be the set of all individuals that are born and $U_t$ the set of those born before time $t$.

FIRST-BIRTHS AND HEIGHTS. Asking about the height of a tree reduces to asking how big the tree needs to be for the first node to appears in some fixed generation. For the branching random walk, the related question is slightly different, and is just to ask how much time does one need to wait to see an individual in the $n$-th generation. This leads to introduce $B_n = \inf\{z_r^{(n)}, r \geq 1\}$ the time of the first birth in the $n$-th generation. This question is at the origin of the branching process techniques to find the heights of random trees. Writing $H_t$ for the number of generations in $U_t$, the main link is

$$H_t = \sup\{n : B_t \leq t\} \quad \text{and} \quad B_n = \inf\{t : H_t \geq n\}.$$

It has been addressed by Hammersley (1974) when all the displacements from a parent are a.s. identical, i.e., for the Bellman–Harris process (Athreya and Ney, 1972). Kingman (1975) dealt with the case where $Z$ is concentrated on $[0,\infty)$ and Biggins (1976, 1977) treated its more general form. To state the theorem, we need to introduce more notations. The results are based on large deviations, but the settings used by the authors of the cited papers are slightly different from ours, and we may use notations that are close to the historical ones.

A LAW OF LARGE NUMBERS. Let $m$ be the Laplace–Stieltjes transform of $Z$, defined by, for $\theta \in \mathbb{R}$,

$$m(\theta) = \mathbf{E}\left[\sum_{r\geq 1} \exp\left(-\theta z_r^{(1)}\right)\right] = \mathbf{E}\left[\int_{-\infty}^{\infty} e^{-\theta t} dZ(t)\right].$$

The function $m$ is very close to our moment generating function, it just handles *all* the children in the mean time instead of taking one at random (which, in general, cannot be as we do it since their number may be infinite). We assume that $m(\theta) < \infty$ for some $\theta > 0$. This implies in particular that $F(t) = \mathbf{E}Z(-\infty, t) < \infty$ for all $t$ and one can then write

$$m(\theta) = \int_{-\infty}^{\infty} e^{-\theta t} dF(t).$$

One then defines the increasing function $\mu$, for $a \in \mathbb{R}$, by

$$\mu(a) = \inf\left\{e^{\theta a} m(\theta) : \theta \geq 0\right\}.$$

This is the equivalent of our Cramér function in multiplicative settings. One can then state the following law of large numbers for $B_n$.

**Theorem 3.6** (Biggins 1976, 1977). *Let $\gamma = \inf\{a : \mu(a) > 1\}$. Let $S$ be the event that the process survives. Then, almost surely on $S$*

$$\frac{B_n}{n} \xrightarrow[n\to\infty]{} \gamma.$$

Theorem 3.6 is the result that Devroye (1986) used in his seminal paper on the height of binary search trees. In our opinion, its great value lies mostly in its generality and the breadth of its applications. We shall pursue similar goals in this thesis, that is, trying to find results that are widely applicable.

FURTHER RESULTS. More precise theorems have been proved about the minimal displacements of a branching random walk. Until recently, they were restricted to very special cases. Precise results about the behavior of $B_n$ include the theorem by Bramson (1978) on the branching Brownian motion, Durrett (1983), and Bachmann (2000). Addario-Berry (2006) and Addario-Berry and Reed (2006) have proved that for a fairly general class of branching random walks, there exists $\gamma$ and $\beta$ such that

$$\mathbf{E}B_n = \gamma n + \beta \log n + O(1).$$

See also the related work of Chauvin and Drmota (2007) on the heights of $m$-ary search trees, and the recent manuscripts by Bramson and Zeitouni (2006) and Hu and Shi (2007).

### 3.2.2 Continuous time branching random walks

A continuous time version of Theorem 3.6 can be found in Biggins (1995, 1996). The presentation here follows the lines of the latter papers. Partial results had appeared in Biggins (1980). As before, a single ancestor is born at the origin at time 0. For convenience, we label the individuals using their line of descent in the Ulam–Harris way: $xy$ denotes the $y$-th child of an individual $x$. We now let the individuals be characterized by not only their position $p_x$, but also the time $\sigma_x$ at which they were born. In these settings, $Z$ is now a spatial point process in $\mathbb{R} \times \mathbb{R}^+$. Each point will correspond to a child. The first coordinate is the deviation from the parent's position and the second the age of the parent when it was born. Let $Z_x$ denote the copy of the point process associated to $x$, with points $\{(z_{xy}, \tau_{xy}), y \geq 1\}$. Then

$$p_{xy} = p_x + z_{xy} \qquad \text{and} \qquad \sigma_{xy} = \sigma_x + \tau_{xy}.$$

The fact that $\tau \geq 0$ ensures that the children are born after their parents. Let $U$ denote the set of individuals that were born.

THE SPREAD OF A BRW. We are now interested in the position of the rightmost individual alive at time $t$, so it is natural to introduce the point process $N_t$ giving the positions at time $t$

$$N_t = \sum_{x \in U} \delta(p_x) \cdot \mathbf{1}[\sigma_x \le t],$$

where $\delta(p)$ is a point at position $p$. The position of the rightmost point at time $t$ is given by

$$B_t = \sup\{p_x : x \in U, \sigma_x \le t\}.$$

We call $\mu$ the intensity measure of $Z$, and $m(\theta, \phi)$ is Laplace–Stieltjes transform:

$$m(\theta, \phi) = \int e^{-\theta z - \phi \tau} \mu(dz, d\tau) = \mathbf{E}\left[\int e^{-\theta z - \phi \tau} Z(dz, d\tau)\right].$$

For supercritical processes, $m(0,0) > 1$, and then one can define

$$\alpha(\theta) = \inf\{\phi : m(\theta, \phi) \le 1\}.$$

Since $m$ is a convex function, $\alpha$ is itself convex. Only $\theta < 0$ has to be considered since we only deal with right extreme points, and hence right tails. The main result of Biggins (1995) is stated as follows:

**Theorem 3.7** (Biggins 1995). *Suppose that $Z$ is supercritical, nonlattice. Assume further that $\mathbf{E}\left[\sup_t e^{-\alpha(\theta)t}\right] < \infty$ and that $\alpha(\theta) < \infty$ for some $\theta < 0$. Then,*

$$\frac{B_t}{t} \to \gamma = \inf\{a : \alpha^\star(a) < 0\},$$

*as $t \to \infty$, where $\alpha^\star$ defined by $\alpha^\star(x) = \inf\{x\theta + \alpha(\theta), \theta < 0\}$ is the concave dual of the convex function $\alpha$.*

**Remarks.** (a) Observe that the functions $m(\cdot,\cdot)$ and $\alpha^\star$ are similar (but not equivalent) to our $\Lambda(\cdot,\cdot)$ and $\Lambda^\star(\cdot,\cdot)$, respectively. The rate function $\Lambda^\star(\cdot,\cdot)$ happens to be convex because of the difference in the definition.

(b) One can also express the limit $\gamma$ in the following way:

$$\gamma = \inf\left\{a : \inf_{\theta<0}\{-\log m(\theta, -a\theta)\} < 0\right\},$$

which may be easier to compute in concrete cases (Biggins, 1980). It seems that $\gamma$ may be interpreted as some kind of slope in a diagram showing $m$. We will see later that it indeed is a slope, but this will be easier to see with our settings.

THE GROWTH OF A BRW. The second parameter that has some interest when comparing branching random walks to random trees is the number of individuals alive at time $t$. Indeed, in most applications, it is of little interest to find the constant $\gamma$ characterizing the limiting behaviour of $B_t$ if one does not know how many individuals are alive in the process at time $t$. This issued has been dealt with by Jagers (1975), Nerman (1981), Cohn (1985) and Biggins (1995, 1996). Let $\xi(t)$ be the number of individuals alive at time $t$, i.e., that were born before $t$:

$$\xi(t) = |\{x \in U : \sigma_x \le t\}| = \sum_{x \in U} \mathbf{1}[\sigma_x \le t].$$

We are after asymptotic estimates for $\xi(t)$. Clearly, one only needs to deal with the time coordinates of the point process. So we may ignore the spatial elements of the point processes. Let $\tilde{\mu}$ be the intensity measure of the point process $\tilde{Z}$, the projection of $Z$ on the time axis, and $\tilde{m}$ its Laplace transform. We have

$$\tilde{m}(\phi) = \int e^{-\phi\tau} \mu(d\tau) = \mathbf{E}\left[\int e^{-\phi\tau} Z(d\tau)\right].$$

The asymptotic size of the branching process depends on the *Malthusian parameter* defined by

$$\alpha = \inf\{\phi : \tilde{m}(\phi) \le 1\}.$$

The next theorem is not the strongest one can state about the size $\xi(t)$ of the tree, but it will be good enough for our purposes. It finds its origins in the much stronger results of Nerman (1981) about the almost sure convergence of a suitably rescaled version of $\xi(t)$ to a nondegenerate random variable.

**Theorem 3.8.** *Let $\xi(t)$ be the size of a supercritical branching process with Malthusian parameter $\alpha$. If the process survives, then*

$$\frac{\log \xi(t)}{\log t} \longrightarrow \alpha$$

*almost surely, as $t \to \infty$.*

**Remarks.** This model of continuous time branching random walk is very close to the *ideal trees* we present in the next chapter. In particular, Theorem 3.7 is in a certain sense the dual of our theorems on the height of ideal trees. The purpose of Chapter 4 is mostly to introduce our notations and interpretations of the asymptotic characterization of the heights.

*Chapter 4*

# An Ideal Model of Random Trees

*In this chapter, we describe an ideal model of random trees. It is tightly related to branching random walks and will help us explain the intuition behind the more general model of Chapter 5. We also prove here the main properties of the geometric interpretation of the height. This is based on part of the work in Broutin et al. (2007) and the early ideas have appeared in Broutin and Devroye (2006).*

> *The tree which moves some to tears of joy is in the eyes of others only a green thing that stands in the way. Some see Nature all ridicule and deformity, and some scarce see Nature at all. But to the eyes of the man of imagination, Nature is Imagination itself.*
> - William Blake, 1799, The Letters

## Contents

## 4.1 Ideal trees: the model

Let $T_\infty$ be an infinite rooted $d$-ary tree (with $d^k$ nodes at every level $k$), and let $r$ be its root. Let $\pi(u)$ be the set of edges on the unique path from a node $u$ up to the root. We assign independently to each node of $T_\infty$ a vector

$$((Z_1, V_1), (Z_2, V_2), \ldots, (Z_d, V_d)),$$

where $V_i \geq 0$, $\sum_{i=1}^d V_i = 1$ and $Z_i \in [-\infty, \infty)$. The components are dependent in a quite arbitrary way, in particular we do not assume any independence between the $V_i$'s and the $Z_i$'s. If an edge $e$ connects $u$ with its $i$-th child, then, for convenience, we define $V_e = V_i$ and $Z_e = Z_i$.

THE SHAPE OF THE TREE. With each node $u \in T_\infty$ we can associate an interval of length $L_u$. We set $L_r = 1$. The children of $u$ have intervals of lengths $L_u \cdot V_1, \ldots, L_u \cdot V_d$ so that the total length $\sum_{i=1}^d L_u V_i = L_u$ is preserved. In this model, the sums of the lengths of the intervals at each level of $T_\infty$ remain 1. The tree thus describes a random sequence of nested partitions. The length of the interval of a node $u$ is $L_u = \prod_{e\in\pi(u)} V_e$. The *ideal tree* with parameter $n$, $T_n$, consists of the nodes $u \in T_\infty$ for which $L_u > 1/n$:

$$T_n = \{u \in T_\infty : L_u > 1/n\}.$$

THE WEIGHTS. The $Z_i$'s represent edge lengths. More specifically, the lengths of the edges connecting $u$ to its children $1, \ldots, d$ are $Z_1, \ldots, Z_d$. In some applications we may have negative values, and in general, the range of each extended random variable $Z_i$ is $[-\infty, \infty)$. We define the *weighted depth* of a node $u \in T_\infty$ by $D_u = \sum_{e\in\pi(u)} Z_e$.

Alternatively, we can see the tree as a birth process. The random vector of interest associated with a node $u$ is then $\mathcal{X}_u = (X_1, \ldots, X_d)$, with $X_i = (Z_i, E_i)$ and $E_i = -\log V_i$, if $V_i > 0$; if $V_i = 0$, we define $E_i = \infty$. The time at which $u$ is born is $B_u = \sum_{e\in\pi(u)} E_e$. In particular, the root is born at time 0. Then, $T_n$ consists of the nodes of $T_\infty$ that are born before time $\log n$. We are interested in the *weighted height*

$H_n$ of $T_n$:

$$H_n = \max\{D_u : u \in T_n\}$$

Since we deal with heights, we may assume without loss of generality that the components $X_1, X_2, \ldots, X_d$ of $\mathcal{X}_u$ are identically distributed. Indeed, randomly permuting them does not affect $H_n$. So, in the sequel, we write $V$, $E$ and $Z$ for the typical distributions of components of $((Z_1, V_1), \ldots, (Z_d, V_d))$ or $\mathcal{X}_u$, and define $X = (Z, E)$.

THE FAIR PORTION OF THE TREE. In general, it is possible that for an edge $e$, $V_e = 0$, $E_e = \infty$ or $Z_e = -\infty$. This ensures that for any $u$ such that $e \in \pi(u)$, $L_u = 0$, $B_u = \infty$, or $D_u = -\infty$, respectively. So, in either case, the node $u$ cannot contribute to the height. Thus we call a node *fair* if $L_u > 0$, $B_u < \infty$ and $D_u > -\infty$. The *fair portion* of the tree is the largest tree rooted at the root of $T_\infty$ consisting of fair nodes only.

We now discuss some constraints on $V$, $E$ and $Z$. Their supports have already been mentioned: $V \in [0,1]$, $E \in [0,\infty]$ and $Z \in [-\infty, \infty)$. Recall that $(Z_i, E_i)$, $1 \le i \le d$ are identically distributed. Also, $\sum_{i=1}^d V_i = 1$, which implies that $\mathbf{E}V = 1/d$. The other conditions, whose relevance is discussed in Section 4.4, are:

(i) $\mathbf{P}\{\exists i : Z_i > -\infty, E_i < \infty\} = 1$: Every fair node $u$ has almost surely (a.s.) at least one fair child.

(ii) $d\mathbf{P}\{Z > -\infty, E < \infty\} > 1$: The fair portion of the tree is not a trivial path.

(iii) $0 \in \mathcal{D}_\Lambda^o$. This implies that $\log \mathbf{E}\left[ e^{\lambda Z} \mid Z > -\infty, E < \infty \right] < \infty$ for some $\lambda > 0$, and in particular we have that $\mathbf{E}[\, Z \mid Z > -\infty, E < \infty\,] < \infty$.

(iv) $\mathbf{E}[\, Z \mid Z > -\infty, E < \infty\,] \ge 0$. Since we study the maximal weighted depth, it seems a natural condition to impose on $Z$.

(v) $\mathbf{E}[\, E \mid Z > -\infty, E < \infty\,] > 0$. This prevents $E$ from being identically 0 on the fair portion of $T_\infty$. In particular, as $E_i = -\log V_i$ and $\sum_i V_i = 1$, this ensures $\mathbf{P}\{E = 0\} < 1/d$.

**Remark.** If $\kappa = \mathbf{P}\{Z > -\infty, E < \infty\} = 1$, then $T_\infty$ is fair. Then, we only need $0 \in \mathcal{D}_\Lambda^o$, $\mathbf{E}Z \geq 0$ and $\mathbf{E}E > 0$.

Under these constraints, the first term of the asymptotic expansion of the weighted height can be characterized by an implicit equation involving large deviation rate functions $\Lambda^\star = \Lambda_X^\star$ (see Chapter 2).

**Theorem 4.1.** *Let $T_n$ be an ideal tree built from $X$ and let $H_n$ be its weighted height. Let $\Lambda^\star = \Lambda_X^\star$. Assume that the conditions* (i) *to* (v) *hold. Then*

$$H_n = c \log n + o(\log n)$$

*in probability, as $n \to \infty$, where $c = \sup\{\alpha/\rho : \Lambda^\star(\alpha, \rho) \leq \log d\}$.*

**Remarks.** (a) Note that, under conditions (i) to (v), $\rho = 0$ is never possible in the supremum defining $c$ (see Lemma 4.1). This is always the case every time we write such a supremum.
(b) If $Z$ and $E$ are independent and do not take infinite values, then $\Lambda^\star(\alpha, \rho) = \Lambda_Z^\star(\alpha) + \Lambda_E^\star(\rho)$, where $\Lambda_Z^\star(\alpha)$ and $\Lambda_E^\star(\rho)$ are defined as the usual Fenchel–Legendre transforms of $\Lambda_Z(\lambda) = \log \mathbf{E}\left[e^{\lambda Z}\right]$ and $\Lambda_E(\mu) = \log \mathbf{E}\left[e^{\mu E}\right]$, respectively. Hence Theorem 4.1 agrees with the result of Broutin and Devroye (2006) which claims that $c$ is the maximal value of $\alpha/\rho$ in $\{\Lambda_Z^\star(\alpha) + \Lambda_E^\star(\rho) \leq \log d\}$. Actually, under their assumptions, the optimal value is attained at a point in $\{\Lambda_Z^\star(\alpha) + \Lambda_E^\star(\rho) = \log d\}$.

The model of ideal trees is clean in the sense that all nodes receive an independent copy of the *same* random vector, and the description of $T_n$ is done in a very natural way from that of $T_\infty$ by pruning the branches. In most cases, the number of nodes of $T_n$ is random, although one expects that it should be close to $n$. However, most concrete applications have random trees of deterministic size. Broutin and Devroye (2006) deal with this by proving that if $E$ is in a specified class of random variables, namely exponentials, the number of nodes of $T_n$ is indeed $n^{1+o(1)}$ in probability. Remarks about what could be a more natural (or useful) definition for the size of $T_n$ can be found in Section 5.6.

## 4.2 Discussions and interpretations

THE DEFINITION OF THE CONSTANT $c$. Before we proceed to proving Theorem 4.1, we shall take it for granted and work on the characterization itself. Indeed, there is much to say about the constant $c$. Recalling about the level sets of $\Lambda^\star$, $\Psi(\ell) = \{(\alpha,\rho) \in \mathbb{R}^2 : \Lambda^\star(\alpha,\rho) \le \ell\}$, we have

$$c = \sup\{\alpha/\rho : \Lambda^\star(\alpha,\rho) \le \log d\} = \sup\{\alpha/\rho : (\alpha,\rho) \in \Psi(\log d)\}.$$

We should also pave the road in preparation of the proof, and gather some alternative definitions for $c$. In the sequel, we let $(Z^c, E^c)$ be distributed as $(Z, E)$ conditional on $\{Z > -\infty, E < \infty\}$. We first argue about the definition of $c$ itself.

**Lemma 4.1.** *Assume* $0 \in \mathcal{D}_\Lambda^o$ *and* $\kappa = \mathbf{P}\{Z > -\infty, E < \infty\} \ge 1/d$.

*(a) If* $(\mathbf{E}Z^c, \mathbf{E}E^c) \in \Psi(\log d) \ne \varnothing$, *then* $c = \sup\{\alpha/\rho : \Lambda^\star(\alpha,\rho) \le \log d\}$ *is well defined.*

*(b) If furthermore,* $\mathbf{E}Z \ge 0$ *then* $c \ge 0$.

*(c) If* $\mathbf{P}\{E = 0\} < 1/d$, *then there exists* $\delta > 0$ *such that*

$$c \le \sup\{\alpha/\rho : \Lambda^\star(\alpha,\rho) \le \log d + \delta\} < \infty.$$

*Proof.* Note that since $0 \in \mathcal{D}_\Lambda^o$, by Lemma 2.2 (c), $(\mathbf{E}Z^c, \mathbf{E}E^c) \in \mathbb{R}^2$.

(a) For any $\lambda, \mu \in \mathbb{R}$, by Jensen's inequality,

$$\Lambda(\lambda,\mu) = \log \mathbf{E}\left[e^{\lambda Z^c + \mu E^c}\right] + \log\kappa \ge \lambda \mathbf{E}Z^c + \mu \mathbf{E}E^c + \log\kappa.$$

It follows that $\lambda\mathbf{E}Z^c + \mu\mathbf{E}E^c - \Lambda(\lambda,\mu) \le -\log\kappa$ and thus, $\Lambda^\star(\mathbf{E}Z^c, \mathbf{E}E^c) \le -\log\kappa$. Since $\kappa \ge 1/d$, $(\mathbf{E}Z^c, \mathbf{E}E^c) \in \{(\alpha,\rho) : \Lambda^\star(\alpha,\rho) \le \log d\} \ne \varnothing$.

(b) If $\mathbf{E}Z^c \ge 0$ we have $c \ge \mathbf{E}Z^c/\mathbf{E}E^c \ge 0$, potentially infinite if $\mathbf{E}E^c = 0$.

(c) For all $\delta > 0$, $c \le \sup\{\alpha/\rho : \Lambda^\star(\alpha,\rho) \le \log d + \delta\}$, so we need only prove that the right-hand side is finite for some $\delta$. Since $\mathbf{P}\{E = 0\} < 1/d$, we can pick $\delta > 0$ such that $\mathbf{P}\{E = 0\} < e^{-\delta}/d$. By Lemma 2.2 (c), $\Lambda^\star$ is a good rate function, and hence the level sets $\Psi(\cdot)$ are compact. As a consequence, it suffices to prove that

$\{\rho = 0\} \cap \Psi(\log d + \delta) = \varnothing$. We show that $\inf_{\alpha\in\mathbb{R}} \liminf_{\rho\downarrow 0} \Lambda^\star(\alpha, \rho) \geq \log d + \delta$, which would prove the claim. For all $\alpha, \rho \in \mathbb{R}$,

$$\Lambda^\star(\alpha,\rho) = \sup_{\lambda,\mu\in\mathbb{R}} \{\lambda\alpha + \mu\rho - \Lambda(\lambda,\mu)\} \geq \sup_{\mu\in\mathbb{R}}\{\mu\rho - \Lambda(0,\mu)\} = \Lambda^\star_{E^c}(\rho) - \log\kappa.$$

Let $q = \mathbf{P}\{E^c = 0\}$. Then,

$$\Lambda_{E^c}(\mu) = \log \mathbf{E}\left[e^{\mu E^c}\right] = \log\left(q + (1-q)\cdot \mathbf{E}\left[\, e^{\mu E^c} \;\middle|\; E^c > 0\,\right]\right).$$

For $\rho_n \downarrow 0$,

$$\begin{aligned}\Lambda^\star_{E^c}(\rho_n) &= \sup_{\mu\in\mathbb{R}}\left\{\mu\rho_n - \log\left(q + (1-q)\mathbf{E}\left[\, e^{\mu E^c} \;\middle|\; E^c > 0\,\right]\right)\right\} \\ &\geq -\sqrt{\rho_n} - \log\left(q + (1-q)\mathbf{E}\left[\, e^{-E^c/\sqrt{\rho_n}} \;\middle|\; E^c > 0\,\right]\right) \to -\log q.\end{aligned}$$

So $\liminf_{n\to\infty} \Lambda^\star_{E^c}(\rho_n) \geq -\log q$. Therefore, for any $\alpha \in \mathbb{R}$,

$$\liminf_{\rho\downarrow 0} \Lambda^\star(\alpha,\rho) \geq -\log \mathbf{P}\{E^c = 0\} - \log\kappa = -\log\mathbf{P}\{E = 0\} > \log d + \delta,$$

and $\inf_\alpha \liminf_{\rho\downarrow 0} \Lambda^\star(\alpha,\rho) \geq \log d + \delta$, which completes the proof. □

A GEOMETRIC INTERPRETATION. Now that we know that $c$ is well-defined and finite under the constraints we have imposed, let us try to characterize it geometrically. Observe that in a diagram where we plot $\alpha$ against $\rho$, $\alpha/\rho$ is the slope of a line with one end at the origin and the other one at $(\alpha, \rho)$. In such a diagram, $\Psi(\log d)$ is a compact set by Lemma 2.2. This set happens to be convex, but this is irrelevant for our point. Then, if one imagines that $\Psi(\log d)$ is embossed, $c$ is the just the slope of the line with a joint at the origin that would be dropped from the vertical. This is illustrated by Figure 4.1. One can also picture of a three-dimensional diagram in which the value of $\Lambda^\star(\alpha,\rho)$ or $I(\alpha,\rho)$ is plotted against $(\alpha,\rho)$. We emphasize this three-dimensional approach since it will be helpful in seeing the parallels between random trees and random tries later. See Figure 4.2.

The following alternate expressions for the constant $c$ will be useful in the proofs, and makes the parallel between $\Lambda^\star(\cdot,\cdot)$ and $I(\cdot,\cdot)$. To understand what is going on in Lemma 4.2, see Figure 4.1.

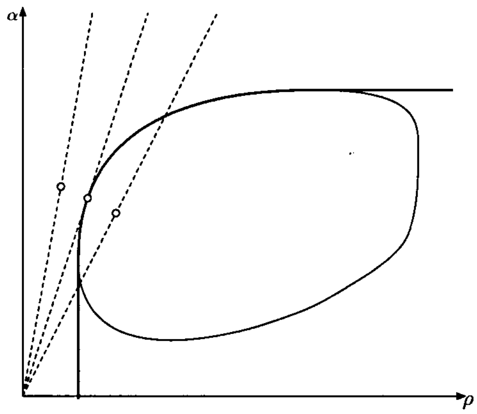


*Figure 4.1: Typical level sets for $\Lambda^*$ and $I$ are shown. The shaded region is the set $\Psi(\log d) = \{\Lambda^*(\alpha, \rho) \leq \log d\}$. The thick line is the border of $\{I(\alpha, \rho) \leq \log d\}$. We also show three points together with the lines of interest. The steepest is used for the upper bound, the most gentle for the lower bound. The intermediate one is the optimal line.*

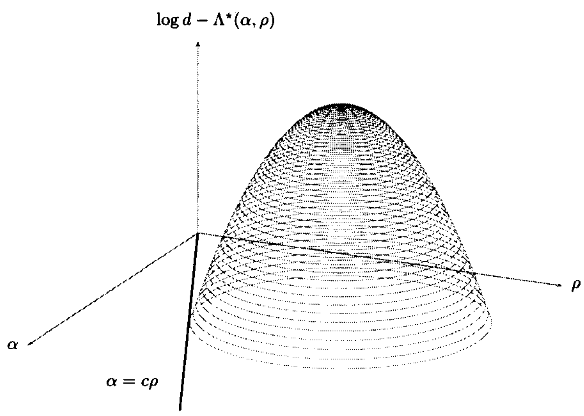


*Figure 4.2: The three-dimensional representation. For reasons that will become clear in the next chapters, we have represented the positive portion of $\log(d) - \Lambda^*(\alpha, \rho)$. The optimal line $\{\alpha = c\rho\}$ in the horizontal plane going through the origin is also shown.*

**Lemma 4.2.** *Suppose that $0 \in \mathcal{D}^o_\Lambda$, $\mathbf{P}\{Z > -\infty, E < \infty\} \geq 1/d$ and $\mathbf{P}\{E = 0\} < 1/d$. Let $c \stackrel{def}{=} \sup\{\alpha/\rho : \Lambda^*(\alpha, \rho) \leq \log d\}$. Then*

*(a) $c = \inf_{\epsilon>0} \sup\{\alpha/\rho : \Lambda^*(\alpha, \rho) \leq \log d + \epsilon\}$.*

*(b) $c = \sup\{\alpha/\rho : I(\alpha, \rho) \leq \log d\}$.*

*(c) $c = \sup\{\alpha/\rho : (\alpha, \rho) \in \Psi(\log d) \cap \mathcal{D}^o_{\Lambda^*}\}$.*

*Proof.* Observe that Lemma 4.1 ensures that $\Psi(\log d) \neq \varnothing$ and that $c$ is well-defined.

(a) Since $\{\Lambda^*(\alpha, \rho) \leq \log d\} \subseteq \{\Lambda^*(\alpha, \rho) \leq \log d + \epsilon\}$ for all $d \geq 1$ and $\epsilon > 0$, it is straightforward that $\sup\{\alpha/\rho : \Lambda^*(\alpha, \rho) \leq \log d\} \leq \inf_{\epsilon>0} \sup\{\alpha/\rho : \Lambda^*(\alpha, \rho) \leq \log d + \epsilon\}$.

For $n \geq 1$, write $c_n \stackrel{\text{def}}{=} \sup\{\alpha/\rho : \Lambda^*(\alpha, \rho) \leq \log d + 1/n\}$. By Lemma 4.1 (c), there exists $n_0$ large enough that $c_n < \infty$ for all $n \geq n_0$. Let $\delta > 0$. For $n \geq n_0$, let $(\alpha_n, \rho_n) \in \Psi(\log d + 1/n)$ be a sequence of points such that $\alpha_n/\rho_n \geq c_n - \delta$. Clearly

$(\alpha_n, \rho_n) \in \Psi(\log d + 1/n_0)$ for all $n \geq n_0$ and since $\Psi(\log d + 1/n_0)$ is compact, there exists a subsequence $\{(\alpha_n, \rho_n), n \geq n_0\}$ that converges to $(\alpha_\infty, \rho_\infty) \in \Psi(\log d + 1/n_0)$ as $n \to \infty$. For each $n \geq n_0$ in the subsequence we have that $\Lambda^\star(\alpha_n, \rho_n) \leq \log d + 1/n$, and since $\Lambda^\star$ is continuous in $\Psi(\log d + 1/n_0)$ (since it is compact), then $\Lambda^\star(\alpha_\infty, \rho_\infty) \leq \log d$. Thus $\alpha_\infty/\rho_\infty \leq \sup\{\alpha/\rho : \Lambda^\star(\alpha, \rho) \leq \log d\}$. Also, since $\alpha_n/\rho_n \geq c_n - \delta$ for all $n \geq n_0$, then $\alpha_\infty/\rho_\infty \geq \lim_{n\to\infty} c_n - \delta = \inf_{\epsilon > 0} \sup\{\alpha/\rho : \Lambda^\star(\alpha, \rho) \leq \log d + \epsilon\} - \delta$. Taking $\delta \to 0$ concludes the proof.

(b) Recall that the rate function $I$ is defined by $I(\alpha, \rho) = \inf\{\Lambda^\star(x, y) : x > \alpha, y < \rho\}$, for $\alpha, \rho \in \mathbb{R}$. We fist show that $\sup\{\alpha/\rho : \Lambda^\star(\alpha, \rho) \leq \log d\} \leq \sup\{\alpha/\rho : I(\alpha, \rho) \leq \log d\}$. For any $\epsilon > 0$, we can pick $\alpha_0 < \alpha$ and $\rho_0 > \rho$ such that $\alpha_0/\rho_0 > \alpha/\rho - \epsilon$. Then $(\alpha_0, \rho_0) \in \Psi_I(\log d)$, implying that $\sup\{\alpha/\rho : \Lambda^\star(\alpha, \rho) \leq \log d\} \leq \sup\{\alpha/\rho : I(\alpha, \rho) \leq \log d\} + \epsilon$. Since $\epsilon$ is arbitrary, $\sup\{\alpha/\rho : \Lambda^\star(\alpha, \rho) \leq \log d\} \leq \sup\{\alpha/\rho : I(\alpha, \rho) \leq \log d\}$.

Next, we show that $\sup\{\alpha/\rho : I(\alpha, \rho) \leq \log d\} \leq \sup\{\alpha/\rho : \Lambda^\star(\alpha, \rho) \leq \log d\}$. Assume that $(\alpha, \rho)$ is accounted in the left-hand side, or $(\alpha, \rho) \in \Psi_I(\log d)$. Then by definition, there exist $(x, y)$ such that $x > \alpha$, $y < \rho$ and $\Lambda^\star(x, y) \leq \log d$. Clearly, $(x, y) \in \Psi(\log d)$ and $x/y > \alpha/\rho$, which proves the claim.

(c) Since $\Lambda^\star$ is finite on $\Psi(\log d)$, we see that $\Psi(\log d)^o \subset \Psi(\log d) \cap \mathcal{D}^o_{\Lambda^\star} \subset \Psi(\log d)$. By Lemma 4.1, $\{\rho = 0\} \cap \Psi(\log d) = \varnothing$, and hence $\alpha/\rho$ is continuous on $\Psi(\log d)$. Accordingly, $\sup\{\alpha/\rho : (\alpha, \rho) \in \Psi(\log d)^o\} = \sup\{\alpha/\rho : (\alpha, \rho) \in \Psi(\log d)\}$. The result follows. □

**Lemma 4.3.** *Assume* $0 \in \mathcal{D}^o_\Lambda$. *Let* $\kappa = \mathbf{P}\{Z > -\infty, E < \infty\}$. *Then* $\Psi(-\log\kappa) \subset \{(\alpha, \rho) : \alpha \leq \mathbf{E}Z^c, \rho \geq \mathbf{E}E^c\}$, *where* $(Z^c, E^c)$ *denotes a random vector distributed as* $(Z, E)$ *conditioned on* $\{Z > -\infty, E < \infty\}$.

*Proof.* We have $\Lambda^\star = \Lambda^\star_{(Z^c,E^c)} - \log\kappa$. So it suffices to prove the claim when $\kappa = 1$, and hence $(Z, E) = (Z^c, E^c)$ almost surely. Assume that $\alpha > \mathbf{E}Z^c$. For any $\rho \in \mathbb{R}$, we have $\Lambda^\star(\alpha, \rho) = \sup_{\lambda,\mu}\{\lambda\alpha + \mu\rho - \Lambda(\lambda, \mu)\} \geq \sup_\lambda\{\lambda\alpha - \Lambda(\lambda, 0)\}$. Since $0 \in \mathcal{D}^o_\Lambda$, $\Lambda$ is differentiable at 0 and $\Lambda(\lambda, 0) = \lambda\mathbf{E}Z^c + o(\lambda)$, as $\lambda \to 0$. As a consequence,

$\lambda\alpha - \Lambda(\lambda, 0) = \lambda(\alpha - \mathbf{E}Z^c) + o(\lambda) \sim \lambda(\alpha - \mathbf{E}Z^c)$ by the assumption. It follows that there exist $\lambda > 0$ such that $\lambda\alpha - \Lambda(\lambda, 0) > 0$, and hence $\Lambda^\star(\alpha, \rho) > 0$, hence proving that $(\alpha, \rho) \notin \Psi(0)$ if $\alpha > \mathbf{E}Z^c$. The case when $\rho < \mathbf{E}E^c$ is treated in a similar way. $\square$

AROUND THE OPTIMAL VALUE. To prove Theorem 4.1, we shall need to show that, for $\epsilon > 0$, $\mathbf{P}\{H_n \geq (c+\epsilon)\log n\} = o(1)$, and $\mathbf{P}\{H_n \geq (c-\epsilon)\log n\} = 1 - o(1)$. In other words, taking for granted the link between these tail probabilities and $\Lambda^\star(\cdot, \cdot)$ and $I(\cdot, \cdot)$, we need some information about the behavior of the curves around $\{\alpha = c\rho\}$. This is why the next lemma is the key to proving the upper and lower bounds of Theorems 4.1.

**Lemma 4.4.** *Assume* $0 \in \mathcal{D}_\Lambda^o$ *and* $\kappa = \mathbf{P}\{Z > -\infty, E < \infty\} > 1/d$. *Let* $c \stackrel{def}{=} \sup\{\alpha/\rho : \Lambda^\star(\alpha, \rho) \leq \log d\}$.

*(a)* $\forall \epsilon > 0$, *there exists* $(\alpha, \rho) \in \mathbb{R}^2$ *such that* $I(\alpha, \rho) < \log d$, *and* $c - \epsilon < \alpha/\rho < c$.

*(b) If* $\mathbf{P}\{E = 0\} < 1/d$, *then, for any* $\epsilon > 0$, $\inf\{\Lambda^\star(x, y) : x/y \geq c + \epsilon\} > \log d$.

*Proof.* Lemma 4.1 ensures that $\Psi(\log d) \neq \varnothing$.

(a) Let $\epsilon > 0$. By definition, we can pick $(\alpha_0, \rho_0)$ such that $\Lambda^\star(\alpha_0, \rho_0) \leq \log d$ and $\alpha_0/\rho_0 > c - \epsilon/2$. Consider the region $\Psi(\log d) \cap \mathcal{B} \cap \mathcal{D}_{\Lambda^\star}^o$ where $\mathcal{B}$ is a non-empty open ball centered at $(\alpha_0, \rho_0)$ for which all points $(x, y) \in \mathcal{B}$ satisfy $x/y > c - \epsilon$. Since $\Lambda^\star$ is convex and $\Lambda^\star(\mathbf{E}Z^c, \mathbf{E}E^c) = -\log\kappa < \log d$, this implies that $\Lambda^\star(x, y) < \log d$ for some $(x, y) \in \Psi(\log d) \cap \mathcal{B} \cap \mathcal{D}_{\Lambda^\star}^o$. Furthermore, we can pick such an $(x, y)$ such that $x/y < c$. Next, pick $(\alpha, \rho) \in \Psi(\log d) \cap \mathcal{B}$ such that $\alpha < x$ and $\rho > y$ (and hence $\alpha/\rho < c$). Since $\Lambda^\star(x, y) < \log d$, then $I(\alpha, \rho) < \log d$, and $c - \epsilon < \alpha/\rho < c$.

(b) Let $\epsilon > 0$ and assume for a contradiction that $\inf\{\Lambda^\star(\alpha, \rho) : \alpha/\rho \geq c + \epsilon\} \leq \log d$. Then, for any $\delta > 0$, there exist $(\alpha, \rho)$ such that $\Lambda^\star(\alpha, \rho) \leq \log d + \delta$ and $\alpha/\rho \geq c + \epsilon$. As a consequence, $\forall\delta > 0$, $\sup\{\alpha/\rho : \Lambda^\star(\alpha, \rho) \leq \log d + \delta\} \geq c + \epsilon$. This implies that $\inf_{\delta>0}\sup\{\alpha/\rho : \Lambda^\star(\alpha, \rho) \leq \log d + \delta\} \geq c + \epsilon$. By by Lemma 4.2 (b), we have that $\inf_{\delta>0}\sup\{x/y : \Lambda^\star(\alpha, \rho) \leq \log d\} = c$, and therefore a contradiction. $\square$

## 4.3 The height of ideal trees

### 4.3.1 The upper bound

We are interested in bounding $\mathbf{P}\{H_n > c' \log n\}$ from above, where $c' = c + \epsilon$, $\epsilon > 0$. We have

$$\mathbf{P}\{H_n > c' \log n\} = \mathbf{P}\{\exists v \in T_n : D_v > c' \log n\}.$$

Let $\mathscr{L}_k$ be the set of nodes $k$ levels away from the root. Then, by the union bound over the levels $k$,

$$\mathbf{P}\{H_n > c' \log n\} \leq \sum_{k \geq 0} \mathbf{P}\{\exists v \in \mathscr{L}_k : D_v > c' \log n, v \in T_n\}.$$

Let $u_k$ be the node in $\mathscr{L}_k$ down the left-most path from the root in $T_\infty$. Using now the union bound on the nodes in generation $k$,

$$\mathbf{P}\{\exists v \in T_n : D_v > c' \log n\} \leq \sum_{k \geq 0} d^k \cdot \mathbf{P}\{D_{u_k} > c' \log n, u_k \in T_n\}. \tag{4.1}$$

Consider now a single term of (4.1), and observe that $\pi(u_k)$ contains $k$ edges:

$$\mathbf{P}\{D_{u_k} > c' \log n, u_k \in T_n\} = \mathbf{P}\left\{ \sum_{e \in \pi(u_k)} Z_e > c' \log n, \sum_{e \in \pi(u_k)} E_e < \log n \right\}.$$

We now distinguish two cases depending on the value of $k$. Let $K \geq 1$ to be chosen later.

THE LOW RANGE. If $k \leq K$, there are few edges in $\pi(u_k)$, and hence it is unlikely that $D_{u_k}$ is large. Since $\mathbf{E}e^{\lambda Z} < \infty$ for some $\lambda > 0$ because $0 \in \mathcal{D}_\Lambda^o$, we have, for this $\lambda$, by Markov's inequality,

$$\mathbf{P}\{D_{u_k} > c' \log n\} \leq \mathbf{P}\left\{e^{\lambda D_{u_k}} \geq e^{\lambda c' \log n}\right\} \leq \frac{\left(\mathbf{E}e^{\lambda Z}\right)^k}{n^{\lambda c'}}.$$

It follows that, for some constant $A = \max\{1, \mathbf{E}e^{\lambda Z}\}$,

$$\sup_{k \leq K} \left\{ \left(\mathbf{E}e^{\lambda Z}\right)^k \right\} \leq A^K,$$

and hence,

$$\sum_{k\le K} d^k \cdot \mathbf{P}\{D_{u_k} > c' \log n, u_k \in T_n\} \le \frac{KA^K}{n^{\lambda c'}}. \tag{4.2}$$

THE DEEP RANGE. We now deal with values of $k$ such that $k \ge K$. Let $\Gamma = \{(x,y) : c'y \le x\}$. Then, We have

$$\begin{aligned} \mathbf{P}\{D_{u_k} > c'\log n, u_k \in T_n\} &\le \mathbf{P}\left\{\left(\sum_{e\in\pi(u_k)} Z_e, \sum_{e\in\pi(u_k)} E_e\right) \in k\cdot\Gamma\right\} \\ &\le \exp\left(-k \inf_{(x,y)\in\Gamma} I(x,y) + o(k)\right), \end{aligned}$$

by Cramér's theorem. We have

$$\inf_{(x,y)\in\Gamma} I(x,y) \ge \inf_{(x,y)\in\Gamma} \Lambda^\star(x,y) \ge \log d + \beta,$$

for some $\beta > 0$, by Lemma 4.4. Therefore, using the lower bound above,

$$\sum_{k\ge K} \mathbf{P}\{\exists v \in \mathscr{L}_k : D_v > c'\log n, v \in T_n\} \le \sum_{k\ge K} d^k e^{-k(\beta+\log d)+o(k)} = O\left(e^{-K\beta/2}\right).$$

FINISHING UP. Plugging the latter bound and (4.2) in (4.1) yields,

$$\mathbf{P}\{H_n > (c+\epsilon)\log n\} \le O\left(e^{-K\beta/2}\right) + O\left(\frac{KB^K}{n^{\lambda c'}}\right).$$

The first term on the right-hand side above can be made as small as we want by picking $K$ large enough. However, since $c \ge 0$, $\lambda c' > 0$. Thus, $K$ being fixed, the second term is made arbitrarily small by letting $n$ go to infinity. This finishes the proof.

### 4.3.2 The lower bound

The proof of the lower bound relies on the construction of a surviving Galton–Watson tree (see Chapter 3). The key ideas are those used in most branching processes proofs of the heights of trees (Devroye, 1986, 1998a) and can be traced back to Biggins (1977).

FINDING DEEP NODES. We start by proving that deep nodes appear in $T_n$ with probability bounded away from zero. Our aim in this section is the following lemma:

**Lemma 4.5.** *For all $\epsilon > 0$, there exists $n_0 \geq 0$ such that*

$$\inf_{n \geq n_0} \mathbf{P} \{\exists u \in T_n : D_u \geq (c - \epsilon) \log n\} \geq 1 - q > 0.$$

*Proof.* If $\epsilon \geq c$, the result is trivial. We assume now that $c - \epsilon > 0$. By Lemma 4.4, there exists $\alpha$ and $\rho$ such that $\alpha/\rho = c' \geq c - \epsilon/2$ and $I(\alpha, \rho) < \log d$. Fix such $\alpha$ and $\rho$. Consider the following branching process defined on $T_\infty$. Let $\ell$ be an arbitrary integer. We call a node $v$ *good* if either it is the root, or $v$ lies $\ell$ levels below a good node $u$ and we have $L_v > L_u \cdot e^{-\ell\rho}$ and $D_v > D_u + \ell\alpha$. The set of good nodes form a branching process. Since $\{\mathcal{X}_u, u \in T_\infty\}$ is a family of i.i.d. random variables, all individuals reproduce independently and in the same way. Therefore, the tree of good nodes we have just built is a Galton–Watson tree. As we have seen in Chapter 3, one determines the behavior of such a process by looking at the average size of the progeny $Y_\ell$ of an individual. By linearity of expectation, writing $\pi(u, v)$ for the set of edges on the unique path from $u$ to $v$,

$$\mathbf{E}Y_\ell = d^\ell \cdot \mathbf{P}\left\{D_v - D_u > \alpha\ell, L_v/L_u > e^{-\rho\ell}\right\},$$

and by definition of $D$ and $L$,

$$\mathbf{E}Y_\ell = d^\ell \cdot \mathbf{P}\left\{\sum_{e \in \pi(u,v)} Z_e > \alpha\ell, \sum_{e \in \pi(u,v)} E_e < \rho\ell\right\}.$$

In the above equation, the right-hand side is exactly the tail probability for a sum of i.i.d. random vectors, as studied in Chapter 2. Since $0 \in \mathcal{D}_\Lambda^o$ by assumption (iii), using Cramér's Theorem (Theorem 2.2), we see that

$$\mathbf{E}Y_\ell = d^\ell e^{-I(\alpha,\rho)\ell + o(\ell)} = e^{\ell \log d - \ell I(\alpha,\rho) + o(\ell)} \xrightarrow[\ell \to \infty]{} \infty,$$

by our choice of $\alpha$ and $\rho$. Thus, there exists $\ell_0$ large enough such that $\mathbf{E}Y_{\ell_0} > 1$. With this choice for $\ell_0$, by Theorem 3.1, the Galton–Watson process survives with positive probability $1-q$. In the case of survival, for every integer $k$, there exists a node $v \in \mathscr{L}_k$ such that $D_v > \alpha\ell_0 k$ and $L_v > e^{-\rho\ell_0 k}$. In particular, for

$$k = \left\lfloor \frac{\log n}{\rho\ell_0} \right\rfloor,$$

and starting the process at the root node of $T_\infty$, there exists $v \in \mathscr{L}_k$ such that $D_v > c' \log n - \alpha\ell_0$ and $L_v > 1/n$. In particular, since $c' \geq c - \epsilon/2$, it happens that $D_v > (c-\epsilon)\log n$, for $n$ large enough. As a consequence, for all $n$ large enough,

$$\mathbf{P}\{\exists v \in T_n : D_v > (c-\epsilon)\log n\} \geq 1-q > 0. \qquad \square$$

Lemma 4.5 above proves the existence of deep nodes in $T_n$ with positive probability. We intend to show that such deep nodes appear in $T_n$ with probability $1-o(1)$ as $n \to \infty$. We want to use a standard boosting argument, and run multiple independent copies of the process described in the proof of Lemma 4.5 by starting at $\mathscr{L}_t$. However, not all $d^t$ such nodes are suitable as starting individuals.

THE NICE PORTION OF THE TREE. A good starting individual $u$ must at least be fair, i.e., satisfy $D_u > -\infty$ and $B_u < \infty$. Because $E$ can take the value $\infty$ or $Z$ the value $-\infty$ with positive probability, we cannot ensure that all nodes at level $t \geq 1$ in $T_\infty$ are fair. However, under our assumptions, enough of them are. We prove this using a second branching process argument.

Let $a, b$ be arbitrary constants to be chosen later. Let a node $u \in T_\infty$ be called *nice* if every edge $e \in \pi(u)$ satisfies $Z_e > a$, and $E_e < b$. Let $R_t$ denote the number of nice nodes at level $t$. Again, $\{R_t, t \geq 0\}$ is a Galton–Watson process. By assumption (ii), $\mathbf{P}\{\exists i : Z_i > -\infty, E_i < \infty\} = 1$. Hence, $\mathbf{P}\{\exists i : Z_i > a, E_i < b\} \to 1$, as $a \to -\infty$ and $b \to \infty$. Also, by (i), $\mathbf{P}\{Z > -\infty, E < \infty\} > 1/d$. Thus, there exist $\delta > 0$, $a_0$ and $b_0$ such that for all $a \leq a_0$ and $b \geq b_0$, $\mathbf{P}\{Z > a, E < b\} > 1/d + \delta$. Therefore, for all $a \leq a_0$ and $b \geq b_0$, we have $\mathbf{E}R_t \geq 1 + d\delta$. By Theorem 3.5, the process survives with probability at least $1-q'$, and $q' = q'(a,b)$ can be made as small as we want by choice of $a$ and $b$.

If the process survives, then by Theorem 3.2, $R_t \to \infty$ as $t \to \infty$ with probability one. Accordingly, for any integer $r \geq 1$, conditioning on survival, there exists $t_0$ such that

$$\mathbf{P}\{R_{t_0} \leq r \mid R_t > 0, \forall t \geq 0\} \leq 1/r. \tag{4.3}$$

To put it differently, we can control the probability that the the number of nice nodes in $\mathscr{L}_t$ is small.

BOOSTING THE SURVIVAL PROBABILITY. Let $\epsilon > 0$. Consider $\{T_\infty(v_i), 1 \leq i \leq R_{t_0}\}$, the family the subtrees of $T_\infty$ rooted at the nice nodes $\{v_i, 1 \leq i \leq R_{t_0}\}$ in $\mathscr{L}_{t_0}$. By Lemma 4.5, there is $n_0$ large enough such that, for all $i$, $1 \leq i \leq R_{t_0}$, we have

$$\mathbf{P}\left\{\exists u \in T_\infty(v_i) : D_u - D_{v_i} > \left(c - \frac{\epsilon}{2}\right)\log m, B_u - B_{v_i} < \log m\right\} \geq 1 - q, \tag{4.4}$$

for all $m \geq n_0$. Recall that $v_i$ is nice, and hence $D_{v_i} \geq at_0$ and $B_{v_i} \leq bt_0$, for all $1 \leq i \leq R_{t_0}$. Let $n$ be large enough. Let $m$ be such that $\log m = \log n - bt_0$. If we can find a node $u$ as in (4.4) in any of $T_\infty(v_i)$, $1 \leq i \leq R_{t_0}$, then $B_u < \log m + B_{v_i} < \log n$ and

$$D_u \;\geq\; at_0 - \left(c - \frac{\epsilon}{2}\right)t_0 b + \left(c - \frac{\epsilon}{2}\right)\log n \;\geq\; (c - \epsilon)\log n$$

for $n$ large enough. Therefore, $u \in T_n$ and $H_n \geq D_u \geq (c - \epsilon)\log n$. Also, for $n \geq n_0$,

$$\begin{aligned}
\mathbf{P}\{H_n \leq (c-\epsilon)\log n\} &\leq \mathbf{P}\{R_{t_0} \leq r\} + \mathbf{P}\{H_n \leq (c-\epsilon)\log n \mid R_{t_0} \geq r\} \\
&\leq \mathbf{P}\{R_{t_0} = 0\} + \mathbf{P}\{1 \leq R_{t_0} \leq r\} + q^r,
\end{aligned}$$

by independence of $T_\infty(v_i)$, $1 \leq i \leq R_{t_0}$. Therefore,

$$\mathbf{P}\{H_n \leq (c-\epsilon)\log n\} \leq q' + \frac{1}{r} + q^r.$$

This can be made as small as we want by choice of $q' = q'(a, b)$ (via $a$ and $b$) and $r$, independently of $n$. This completes the proof of the lower bound of Theorem 4.1.

## 4.4 Special cases

We have imposed constraints (i) to (v) on our random trees. We now discuss some of the cases when they don't hold, and see what happens with the weighted height $H_n$.

THE FAIR PORTION OF THE TREE IS FINITE. Intuitively, if the fair portion of the tree is finite, then, the unweighted height is finite, and the weighted height should be finite as well. This is formalized in the following lemma.

**Lemma 4.6.** *Suppose that* $0 \in \mathcal{D}_\Lambda^o$. *Assume that* $\mathbf{P}\{Z > -\infty, E < \infty\} < 1/d$ *or that* $\mathbf{P}\{Z > -\infty, E < \infty\} = 1/d$ *and* $\mathbf{P}\{\exists i : Z_i > -\infty, E_i < \infty\} < 1$. *Then* $H_n = O(1)$ *in probability, as* $n \to \infty$.

*Proof.* The tree consisting of fair nodes, $\{u \in T_\infty : D_u > -\infty, B_u < \infty\}$, is disbributed as a Galton–Watson process. The expected number of children of an individual is $d\mathbf{P}\{Z > -\infty, E < \infty\}$. So, either the process is subcritical or it is critical and not degenerated to a path. In both cases, the tree of fair nodes is almost surely finite. If follows easily that the weighted height is bounded in probability since $\mathbf{E}[\, Z \mid Z > -\infty, E < \infty\,] < \infty$. □

THE FAIR PORTION OF THE TREE IS A PATH. This is the degenerate case where where, in essence, we have a random walk instead of a branching random walk. The weighted height can be characterized by following the lines of the proof of Theorem 4.1 in this special case.

**Lemma 4.7.** *Write* $(Z^c, E^c)$ *for the distribution of* $(Z, E)$ *conditioned on* $\{Z > -\infty, E < \infty\}$, *and assume that we have* $\mathbf{P}\{\exists i : E_i < \infty, Z_i > -\infty\} = 1$ *and* $\kappa = \mathbf{P}\{E < \infty, Z > -\infty\} = 1/d$. *Then, as* $n \to \infty$

$$H_n = \left(\frac{\mathbf{E}Z^c}{\mathbf{E}E^c} + o(1)\right) \cdot \log n \quad \textit{in probability.}$$

*Proof.* Recall that a node $u$ is fair if $D_u > -\infty$ and $B_u < \infty$. The tree consisting of fair nodes is distributed as a Galton–Watson tree. The expected number of children of an individual is $d\mathbf{P}\{Z > -\infty, E < \infty\} = 1$, hence the process is critical. Also, since $\mathbf{P}\{\exists i : E_i < \infty, Z_i > -\infty\} = 1$, there is a.s. a fair child, and the tree consists of a single infinite path. Let this path be $\{v_i, i \geq 1\}$, characterized by the pairs $\{(Z_i^c, E_i^c), i \geq 1\}$, from the root down. So, for any $\alpha$,

$$\mathbf{P}\{H_n \geq \alpha\} \geq \sup_k \mathbf{P}\left\{\sum_{i=1}^k Z_i^c \geq \alpha, \sum_{i=1}^k E_i^c < \log n\right\}.$$

Set $\epsilon > 0$, $\alpha = (\mathbf{E}Z^c - \epsilon)\log n$, and

$$k = \left\lfloor \frac{\log n}{\mathbf{E}E^c + \epsilon} \right\rfloor .$$

Then, by the law of large numbers, $\mathbf{P}\{H_n \geq (\mathbf{E}Z^c - \epsilon)\log n\} \to 1$, as $n \to \infty$.

It remains to find a matching upper bound. By analogy with the general case, let $c = \mathbf{E}Z^c/\mathbf{E}E^c \geq 0$. Let $K \geq 1$. On the one hand, the nodes with $D_u \to \infty$ cannot lie at constant distance from the root:

$$\begin{aligned} \mathbf{P}\{\exists k \leq K : D_{v_k} > (c+\epsilon)\log n\} &\leq K \sup_{k\leq K} \mathbf{P}\left\{\sum_{i=1}^{k} Z_i^c > (c+\epsilon)\log n\right\} \\ &\leq K \sup_{k\leq K} \frac{k\mathbf{E}Z^c}{(c+\epsilon)\log n} = O\left(\frac{K^2}{\log n}\right), \end{aligned} \tag{4.5}$$

by Markov's inequality. On the other hand, for the nodes $v_k$ with $k \geq K$, by Cramér's theorem (Theorem 2.2),

$$\begin{aligned} \mathbf{P}\{D_{v_k} > (c+\epsilon)\log n\} &= \mathbf{P}\left\{\sum_{i=1}^{k} Z_i^c > (c+\epsilon)\log n, \sum_{i=1}^{k} E_i^c < \log n\right\} \\ &\leq \exp\left(-k \inf_{x\geq(c+\epsilon)y} \Lambda^\star(x,y) - k\log d\right), \end{aligned}$$

since the rate functions for $(Z^c, E^c)$ and $(Z, E)$ are translate from one another by $\log d$. We can define $c$ in a similar fashion as used in Theorem 4.1. Indeed, by Lemma 4.3, $\Psi(\log d) = \Psi(-\log \kappa) \subset \{(\alpha,\rho) : \alpha \leq \mathbf{E}Z^c, \rho \geq \mathbf{E}E^c\}$, and hence $c = \sup\{\alpha/\rho : \Lambda^\star(\alpha,\rho) \leq -\log\kappa\} = \sup\{\alpha/\rho : \Lambda^\star(\alpha,\rho) \leq \log d\}$. Thus by an argument similar to that of Lemma 4.4, there exists $\beta > 0$ independent of $k$ or $n$ such that $\inf\{\Lambda^\star(x,y) : x \geq (c+\epsilon)y\} > \log d + \beta$. Then, by the union bound,

$$\mathbf{P}\{\exists k \geq K : D_{v_k} > (c+\epsilon)\log n\} \leq \sum_{k\geq K} e^{-\beta k} = O\left(e^{-\beta K}\right). \tag{4.6}$$

Putting (4.5) and (4.6) together, and since $T_n \subset \{v_k, k \geq 0\}$ a.s., we obtain

$$\mathbf{P}\{\exists u \in T_n, D_u > (c+\epsilon)\log n\} \leq O\left(\frac{K^2}{\log n}\right) + O\left(e^{-\beta K/2}\right).$$

First picking $K$ large enough, and then letting $n$ tend to infinity proves the upper bound. $\square$

POSITIVE HEIGHTS AND NEGATIVE WEIGHTS. Finally, we conclude this section with a remark on the sign of $c$. One could think that when the weights are mostly negative, $c$ should be negative as well. This is *not* the case, as shown by the following example.

**Lemma 4.8.** *There exist independent random variables $Z$ and $E$ such that*

$$\mathbf{E}[\, Z \mid Z > -\infty, E < \infty\,] < 0 \quad \textit{and} \quad \mathbf{P}\{Z > -\infty, E < \infty\} > 1/d,$$

*and yet $H_n = \Omega(\log n)$ in probability.*

*Proof.* Consider a binary tree. Let $(V_1, V_2) = (1/2, 1/2)$. Then $T_n$ is a complete binary tree with $\lfloor \log_2 n \rfloor$ levels. Now, let $(Z_1, Z_2) = (-2, 1)$. Clearly, with the symmetrized random variables $(Z, E)$, $\mathbf{P}\{Z > -\infty, E < \infty\} = 1$ and $\mathbf{E}[\, Z \mid Z > -\infty, E < \infty\,] = \mathbf{E}Z = -1 < 0$. However, there exists a path from the root to a leaf of $T_n$ not contaning any negative $Z$, and hence $H_n = \lfloor \log_2 n \rfloor$. □

## 4.5 The effective size of a tree

In some applications, one wants to express the height of the tree in terms of the number of significant nodes. Only the fair portion of the tree is significant for the height, and we shall define the *effective size* $\#T_n$ of $T_n$ as the size of its fair portion:

$$\#T_n = |\{u \in T_n : D_u > -\infty\}|.$$

When $\mathbf{P}\{Z = -\infty\} = 0$, the effective size is just the number of nodes $|T_n|$. The only difference between the height $H_n$ and that of a tree of effective size $n$ is essentially a scaling factor.

**Theorem 4.2.** *Assume (i) to (v) hold. Let $T_n$ be an ideal tree of (random) effective size $s_n = \#T_n$. Then, its height satisfies $H_n = \frac{c}{\gamma}\log s_n + o(\log s_n)$ in probability, as $n \to \infty$, where $c = \sup\{\alpha/\rho : \Lambda^\star(\alpha, \rho) \le \log d\}$ and $\gamma = -\sup\{\phi : \Lambda_Y(\phi) \le -\log d\}$, with $Y = E + \infty \cdot \mathbf{1}[Z = -\infty]$.*

Theorem 4.2 follows easily from the following estimation of the effective size of $T_n$. It is just a transcription of Theorem 3.8 in our notation.

**Lemma 4.9.** *Assume (i) to (v) hold. Let $T_n$ be an ideal tree build from the random vector $(Z, E)$. Let $\gamma = -\sup\{\phi : \Lambda_Y(\phi) \leq -\log d\}$. Then, as $n \to \infty$, $\log \#T_n \sim \gamma \log n$ in probability.*

*Proof.* The effect of $Z = -\infty$ is to cut down a subtree. We introduce a modified time random variable $Y$ producing the same effect: $Y = E + \infty \cdot \mathbf{1}[Z = -\infty]$. Because the proofs rely on the renewal theorem, Biggins (1996) assumes the distributions are nonlattice. However, this is only an issue due to the proof technique, and the theorems can be proved true in the lattice case as well (Nerman, 1981; Biggins, 1996). Theorem 3.8 can be used without modification, provided we translate it to our setting. We use the cumulant generating function $\Lambda_Y$ defined by

$$\Lambda_Y(\phi) = \log \mathbf{E}\left[e^{\phi Y}\right] + \log \mathbf{P}\{Y < \infty\},$$

for $\phi \in \mathbb{R}$. The *Malthusian* parameter

$$\gamma = -\sup\{\phi : \Lambda_Y(\phi) \leq -\log d\}$$

is the quantity of interest. Here, it turns out that, for all $\phi$,

$$\Lambda_Y(\phi) = \log \mathbf{P}\{Z > -\infty, E < \infty\} + \log \mathbf{E}\left[\, e^{\phi E} \;\middle|\; Z > -\infty, E < \infty \right].$$

Also, $\Lambda_Y(0) = \log \mathbf{P}\{Z > -\infty, E < \infty\} > -\log d$ by assumption. Hence, $\gamma > 0$ (which just means that the process is supercritical). We clearly have that $\sup_t e^{-\gamma t} < \infty$ which implies by Theorem 3.8 that $\log \#T_n \sim \gamma \log n$ on the surviving set. However, by (i) the process survives with probability 1. As a consequence, we have $\log \#T_n \sim \gamma \log n$ a.s. and thus in probability. □

**Remark.** Lemma 4.9 can also be proved using properties of recursive equations in distribution and the contraction method (see Rösler, 1992; Rachev and Rüschendorf, 1995; Rösler and Rüschendorf, 2001).

*Proof of Theorem 4.2.* The proof is straightforward. By Theorem 4.1, $H_n = c \log n + o(\log n)$ in probability as $n \to \infty$. Also, by Lemma 4.9, $s_n = \#T_n \sim \gamma \log n$ in probability, as $n \to \infty$. Now,

$$\frac{H_n}{\log n} \to c \qquad \text{and} \qquad \frac{\log n}{\log s_n} \to \frac{1}{\gamma}$$

in probability, as $n \to \infty$. Therefore, the product converges as well. This can be proved formally but tediously using the following characterization of convergence in probability: every subsequence contains a further subsequence that converges almost surely, or see Exercise 20.20 on p. 272 of Billingsley (1995). $\square$

Chapter 5

# Weighted height of random trees

*We introduce a general model model of weighted random trees based on the ideal trees of Chapter 4. This models permits to obtain the height of pebbled tries, pebbled ternary search tries, d-ary pyramids, and to study geometric properties of partitions generated by k-d trees. The chapter is based on Broutin et al. (2007) and uses earlier ideas of Broutin and Devroye (2006) and Broutin et al. (2006).*

*Down with bushes! Hail to the trees!*
– Common Sense

## Contents

## 5.1 Introduction

The model is largely inspired by that of ideal trees presented in the previous chapter. Again, every node is associated with two random vectors, $\mathcal{Z}$, describing the lengths of the edges to the children, and $\mathcal{V}$, describing the size of the subtrees of the children. However, for the model to be useful and directly applicable to a large number of random tree models, we generalize earlier results in two ways. First, we introduce the notion that only the limiting vectors (as the size of a subtree grows) are relevant. This idea has been used by Broutin et al. (2006) in the unweighted settings. Second, we allow the two random vectors $\mathcal{Z}$ and $\mathcal{V}$ to be dependent. We prove that under some mild conditions on the random vectors, the height of a random tree of size $n$ is asymptotic to $c \log n$ in probability. We characterize $c$ uniquely as the only solution of an (often implicit) equation involving large deviation rate functions, as in the case of ideal trees of Chapter 4.

## 5.2 A model of random trees

Weighted random trees can be constructed using a variety of methods, also called embeddings. An embedding emphasizes an underlying structure consisting of independent random variables. The model that we propose describes one embedding. It is of such generality that many important brands of random trees can be captured by it. Examples follow at the end of this chapter.

DIFFERENT TYPES OF NODES. Consider a family $\{\mathcal{X}^m, m \geq 0\}$ of random vectors, where $\mathcal{X}^m = ((Z_1^m, E_1^m), \ldots, (Z_d^m, E_d^m))$. Assume that for all $m$, and $1 \leq i \leq d$, $m \exp(-E_i^m)$ is almost surely integer-valued, and $E_i^m \geq 0$. Assign independently a copy of the sequence $\{\mathcal{X}^m, m \geq 0\}$ to each one of the nodes of an infinite $d$-ary tree $T_\infty$. The different elements of the sequence $\{\mathcal{X}^m, m \geq 0\}$ allow to describe different behavior for the nodes. In a sense, we have different types of nodes, one for each natural integer.

BUILDING RANDOM TREE ON $n$ ITEMS. Given an integer $n$ and the copies of $\{\mathcal{X}^m, m \geq 0\}$, we build a sequence $\{(D_u, B_u), u \in T_\infty\}$ of weighted depths and birth times of the nodes of $T_\infty$. Observe that although the dependence is not explicitly written, $\{(D_u, B_u), u \in T_\infty\}$ depends on $n$. The construction is made easier by using the auxiliary sequence $\{N_u, u \in T_\infty\}$, where $N_u$ is the *cardinality* of a node $u$, that is the number of items in its subtree. Let $n \geq 0$ and consider $((Z_1^n, E_1^n), \ldots, (Z_d^n, E_d^n))$, the copy of $\mathcal{X}^n$ at the root of $T_\infty$. The children $u_1, \ldots, u_d$ of the root are assigned cardinalities $N_{u_i} = n \exp(-E_i^n) \in \mathbb{N}$, $1 \leq i \leq d$. Given the values of $N_{u_1}, \ldots, N_{u_d}$, the sequences $\{N_v : v \in T_\infty(u_i)\}$, $1 \leq i \leq d$, describing the trees rooted at $u_i$, $1 \leq i \leq d$ are recursively built in the same way, unless $1 \leq N_{u_i} \leq b$ or $N_{u_i} = 0$. Here $b$ is the number of items that a node can contain.

Using $\{N_u, u \in T_\infty\}$, and the copies of $\{\mathcal{X}^m, m \geq 0\}$, we now assign random variables $(Z_e, E_e)$ to the edges of $T_\infty$. Let $e$ be the $i$-th edge out of a node $u \in T_\infty$. We set

$$Z_e = Z_i^{N_u} \quad \text{and} \quad E_e = E_i^{N_u}.$$

Recall that $\pi(u)$ denotes the set of edges on the path from $u$ up to the root in $T_\infty$. As for the case of ideal trees, we define the weighted depth of a node $u$, $D_u = \sum_{e \in \pi(u)} Z_e$ and the birth time of a node $u$, $B_u = \sum_{e \in \pi(u)} E_e$. This finishes the construction of

$\{(D_u, B_u), u \in T_\infty\}$ which fully describes our random weighted tree. Then we have

$$T_n \stackrel{\text{def}}{=} \{u \in T_\infty : N_u > 0\} = \{u \in T_\infty : B_u < \log n\}.$$

We are interested in the weighted height $H_n = \max\{D_u : u \in T_n\}$ of the random tree $T_n$. Again, it is sufficient to consider the trees for which the components of $\mathcal{X}^n$ are identically distributed. We have the following conditions:

- PERMUTATION INVARIANCE. For any integer $n$, and any permutation $\sigma$, the vector $((Z^n_{\sigma(1)}, E^n_{\sigma(1)}), \ldots, (Z^n_{\sigma(d)}, E^n_{\sigma(d)}))$ is distributed as $((Z^n_1, E^n_1), \ldots, (Z^n_d, E^n_d))$.
- CONVERGENCE. There exists a random vector $\mathcal{X}^\infty$ such that the cumulant generating functions of the *vectors* $\mathcal{X}^n$ and $\mathcal{X}^\infty$ satisfy $\Lambda_{\mathcal{X}^n} \to \Lambda_{\mathcal{X}^\infty} \leq \infty$ everywhere as $n \to \infty$ and $0 \in \mathcal{D}^o_{\Lambda_{\mathcal{X}^\infty}}$.
- BOUNDED HEIGHT. There exists a deterministic function $\psi$ such that for all $n$, $H_n \leq \psi(n)$.

**Remarks.** (a) Observe that since $0 \in \mathcal{D}^o_{\Lambda^\infty}$, $\Lambda_{\mathcal{X}} \to \Lambda_{\mathcal{X}^\infty}$ implies that $\mathcal{X} \to \mathcal{X}^\infty$ in distribution (see Billingsley, 1995, p. 390).
(b) We can slightly relax the constraint that the height be bounded. For instance, subexponential tails for the height would suffice: for all $M \geq 1$, there exists a function $f$ with $f(t)/t \to \infty$ as $t \to \infty$ such that

$$\sup_{n \leq M} e^{f(t)} \mathbf{P}\{H_n \geq t\} \leq 1.$$

Ordinary tries violate this condition, and will be treated separately in Chapter 6.

It is also clear from the construction that:

**Lemma 5.1.** *Let $T_n$ be a random tree as defined above. Then we have*

- CONDITIONAL INDEPENDENCE. *For any node $u$, the $\sigma$-algebras generated by the variables associated with edges in the subtrees rooted at the children $u_1, \ldots, u_d$ are independent, conditioned on the sizes $N_{u_1}, \ldots, N_{u_d}$.*

- Size-dependent distribution. *Conditioning on $N_u = k$, the subtree rooted at $u$, $T_n(u)$, is distributed as $T_k$.*

As in the case of ideal trees of Chapter 4, the height may be characterized using large deviation functions. Actually, it turns out that under these constraints, the heights of $T_n$ and an ideal tree built using the vector $\mathcal{X}^\infty$ are asymptotically comparable in probability. We first recall the assumptions for the model of ideal tree. Let $\Lambda$ be the cumulant generating function associated with a typical (uniformly random) component of $X^\infty = (Z^\infty, E^\infty)$ of $\mathcal{X}^\infty$. Then, we require that

(i) $\mathbf{P}\{\exists i : Z_i^\infty > -\infty, E_i^\infty < \infty\} = 1$.

(ii) $\mathbf{P}\{Z^\infty > -\infty, E^\infty < \infty\} > 1/d$.

(iii) $0 \in \mathcal{D}_\Lambda^o$.

(iv) $\mathbf{E}[\, Z^\infty \mid Z^\infty > -\infty, E^\infty < \infty\,] \geq 0$.

(v) $\mathbf{E}[\, E \mid Z^\infty > -\infty, E^\infty < \infty\,] > 0$.

See Chapter 4 for more information about the conditions above. The main result of this chapter, and indeed this thesis, is the following theorem. Let $\Lambda$ be the generating function of the cumulants of $(Z^\infty, E^\infty)$, and let $\Lambda^\star$ be its convex dual (see Chapter 2).

**Theorem 5.1.** *Let $T_n$ be the random tree defined above and let $H_n$ be its weighted height. Suppose that (i) to (v), together with the above conditions hold for $\mathcal{X}^\infty$. Let $c = \sup\{\alpha/\rho : \Lambda^\star(\alpha,\rho) \leq \log d\}$. Then $H_n = c\log n + o(\log n)$ in probability, as $n \to \infty$.*

The heights of many known trees fall within the scope of Theorem 5.1. These include binary search trees (Devroye, 1986), bounded degree increasing trees (Bergeron et al., 1992; Broutin et al., 2006), random recursive trees (Devroye, 1987; Pittel, 1994), plane-oriented trees (Pittel, 1994), scale-free trees (Pittel, 1994; Barabási and Albert, 1999) pyramids (Mahmoud, 1994; Biggins and Grey, 1997), and most models captured by the less general result of Broutin and Devroye (2006). Many applications are treated in section 5.7.

## 5.3 Relying on ideal trees

In the proof of Theorem 5.1 we approximate depths in random trees by those in a suitable ideal tree. We now introduce the following coupling. By assumption, $\mathcal{X}^n \to \mathcal{X}^\infty$ in distribution as $n \to \infty$, hence by Skorohod's theorem (see, e.g., Billingsley, 1995), we can find a coupling for which the convergence is almost sure. In the following, we let $\mathcal{X}^n$ be the copies of the random variables such that $\mathcal{X}^n \to \mathcal{X}^\infty$ almost surely. If we use copies of this coupled sequence $\{\mathcal{X}^m, m \geq 0\}$ to build the random trees, we obtain a coupled sequence $\{T_n, n \geq 0\}$. Since the convergence of $\mathcal{X}^n$ to $\mathcal{X}^\infty$ is almost sure, each node has a copy of $\mathcal{X}^\infty$ as well. These copies, in turn, define a proper ideal tree with split vector $\mathcal{X}^\infty$. This latter tree is called the *ideal tree associated* with the coupled sequence $\{T_n, n \geq 0\}$.

**Lemma 5.2.** *Consider the coupled sequence of random trees $\{T_n, n \geq 0\}$, and the associated ideal tree. Let $\ell$ be a fixed positive integer. Let $v_1, v_2, \ldots, v_k$ be the set of $k = d^\ell$ nodes in generation $\ell$ of $T_\infty$. Then, as $n \to \infty$,*

$$((D_{v_1}, B_{v_1}), (D_{v_2}, B_{v_2}), \ldots, (D_{v_k}, B_{v_k})) \to ((D_{v_1}^\infty, B_{v_1}^\infty), (D_{v_2}^\infty, B_{v_2}^\infty), \ldots, (D_{v_k}^\infty, B_{v_k}^\infty))$$

*a.s., where $D_{v_i}^\infty$ and $B_{v_i}^\infty$ are the weighted depth and birth time of $v_i$ in an ideal tree built from the limit vector $\mathcal{X}^\infty$.*

*Proof.* Since $\mathcal{X}^n \to \mathcal{X}^\infty$ a.s., each node has an independent copy of the limit as well. These limit random variables are used to define $\{(D_u^\infty, B_u^\infty), u \in T_\infty\}$, which characterizes fully a coupled *ideal* tree. Assume for now that, for all $u \in T_\infty$,

$$(D'_u, B'_u) \xrightarrow[n\to\infty]{} (D_u^\infty, B_u^\infty) \quad \text{almost surely.} \tag{5.1}$$

This implies that

$$((D'_{v_1}, B'_{v_1}), \ldots, (D'_{v_k}, B'_{v_k})) \to ((D_{v_1}^\infty, B_{v_1}^\infty), \ldots, (D_{v_k}^\infty, B_{v_k}^\infty))$$

almost surely, as $n \to \infty$. Therefore, to prove the lemma, it suffices to show that (5.1) holds for all $u \in T_\infty$.

Let $A$ be a set of probability 1 on which, for all $u$, $\mathcal{X}_u^n \to \mathcal{X}_u^\infty$. We prove by induction on the (unweighted) depth that

$$\forall \omega \in A \quad (D'_u(\omega), B'_u(\omega)) \to (D_u^\infty(\omega), B_u^\infty(\omega)).$$

For the sake of simplicity, we drop the $\omega$ and simply write $(D'_u, B'_u)$ and $(D_u^\infty, B_u^\infty)$, remembering that, in fact, these values are *deterministic* and taken at the point $\omega$. If $u$ is the root, then $(D'_u, B'_u) = (0,0) = (D_u^\infty, B_u^\infty)$. Otherwise, $u$ is the $i$-th child of some node $v$. The induction hypothesis tells us that $(D'_v, B'_v) \to (D_v^\infty, B_v^\infty)$ as $n \to \infty$. Let the components of $\mathcal{X}^\infty$ be $(Z_i^\infty, E_i^\infty)$, $1 \le i \le d$. Assume first that $B_v^\infty = \infty$, then $B_u^\infty = B_v^\infty + E_i^\infty(v) = \infty$. As $B'_u \ge B'_v$, it follows that $B'_u \to B_u^\infty$. If $B_v^\infty < \infty$, we have $N'_v = n\exp(-B'_v) \sim n\exp(-B_v^\infty) \to \infty$ as $n \to \infty$. As a consequence,

$$\begin{aligned} D'_u &= D'_v + Z_i^{N'_v}(u) \xrightarrow[n\to\infty]{} D_v^\infty + Z_i^\infty(v) = D_u^\infty, \quad \text{and} \\ B'_u &= B'_v + E_i^{N'_v}(v) \xrightarrow[n\to\infty]{} B_v^\infty + E_i^\infty(v) = B_u^\infty. \end{aligned}$$

Therefore, $(D'_u, B'_u) \to (D_u^\infty, B_u^\infty)$, as $n \to \infty$, which completes the proof. □

**Important remark.** Proving Theorem 5.1 amounts to showing that a property holds in probability. As a consequence, we can use the coupled sequence of trees we have just described. In the remaining of the chapter, the trees we consider are always taken from this coupled sequence. In particular, there always exists a coupled ideal tree to rely on, and it does make sense to condition on events happening on this ideal tree to study random variables in $T_n$. We let $Z^\infty$, $E^\infty$, $D^\infty$, and $B^\infty$ be the variables associated with the coupled ideal tree, so for a node $u \in T_\infty$ the variables of interest in the ideal tree are

$$D_u^\infty = \sum_{e \in \pi(u)} Z_e^\infty \qquad \text{and} \qquad B_u^\infty = \sum_{e\in\pi(u)} E_e.$$

## 5.4 The upper bound

Let $\Lambda^n$ denote de cumulant generating function of a typical component $(Z^n, E^n)$ of $\mathcal{X}^n$. Let $\mathscr{L}_k$ be the set of nodes $k$ levels away from the root in $T_\infty$. Let $u_k$ be the left-most node in $\mathscr{L}_k$. We introduce the event $F_k$ defined by

$$F_k \stackrel{\text{def}}{=} \{Z_e > -\infty, E_e < \infty, \forall e \in \pi(u_k)\}.$$

The upper bound is based on the Gärtner–Ellis theorem (Theorem 2.4). The following result proves that the conditions for its application hold, with the event $A_M$ being $N_u \geq M$.

**Lemma 5.3.** *Let $\lambda, \mu \in \mathbb{R}$. For any $\delta > 0$, there exists $M$ large enough that*

$$\sup_{n,k} \left\{ \frac{1}{k} \log \mathbf{E}\left[\, \mathbf{1}[F_k, N_{u_k} \geq M] \cdot \exp\left(\lambda D_{u_k} + \mu B_{u_k}\right) \mid N_{u_0} = n \right] \right\} \leq \Lambda(\lambda, \mu) + \delta.$$

*Proof.* In order to improve the readability of the equations, and for the course of this proof only, let us reindex the random vectors $(Z_e, E_e)$ on the left-most path to $T_\infty$ as $\{(Z_i, E_i), i \geq 1\}$, where the indices increase with the distance from the root. In the same spirit, for $i \geq 0$, write $N_i$, $D_i$ and $B_i$ for $N_{u_i}$, $D_{u_i}$ and $B_{u_i}$, respectively. If $n < M$, we clearly have $\mathbf{1}[N_k \geq M] = 0$ and the result holds. With our new notations, $D_k = \sum_{i=1}^k Z_i$ and $B_k = \sum_{i=1}^k E_i$, so proving the results reduces to bounding

$$\begin{aligned} C &\stackrel{\text{def}}{=} \mathbf{E}\left[\, \mathbf{1}[F_k, N_k \geq M] \cdot e^{\lambda D_k + \mu B_k} \mid N_0 \right] \\ &= \mathbf{E}\left[\, \mathbf{1}[F_k, N_k \geq M] \cdot e^{\sum_{i=1}^k \lambda Z_i + \mu E_i} \;\middle|\; N_0 \right]. \end{aligned}$$

The random vectors $(Z_i, E_i)$ are not independent. However, by conditioning on $N_1$,

$$C = \mathbf{E}\left[\, \mathbf{E}\left[\, \mathbf{1}[F_k, N_k \geq M] \cdot e^{\sum_{i=1}^k \lambda Z_i + \mu E_i} \;\middle|\; N_1 \right] \;\middle|\; N_0 \right].$$

Let $F_k^2$ be the event that $\{Z_i, E_i \in \mathbb{R}, 2 \leq i \leq k\}$. Then, given $N_0$ and $N_1$, the random variables $\mathbf{1}[F_k^2, N_k \geq M] \exp(\sum_{i=2}^k \lambda Z_i + \mu E_i)$ and $\mathbf{1}[F_1] \exp(\lambda Z_1 + \mu E_1)$ are independent. Hence

$$C \leq \mathbf{E}\left[ \underbrace{\mathbf{E}\left[\, \mathbf{1}[F_k^2, N_k \geq M] \cdot e^{\sum_{i=2}^k \lambda Z_i + \mu E_i} \;\middle|\; N_1 \right]}_{I} \cdot \mathbf{E}\left[\, \mathbf{1}[F_1^2] \cdot e^{\lambda Z_1 + \mu E_1} \mid N_1 \right] \;\middle|\; N_0 \right],$$

where we used $\mathbf{1}[N_1 \geq M] \leq 1$ in the second factor. The first factor can be bounded by

$$I \leq \sup_{m \geq M} \mathbf{E}\left[ \mathbf{1}[F_k^2, N_k \geq M] \cdot e^{\sum_{i=2}^{k} \lambda Z_i + \mu E_i} \;\middle|\; N_1 = m \right],$$

which is independent of $N_1$ and $N_0$. Let $\delta > 0$ and let $M$ be large enough that for all $m \geq M$, $\Lambda^m(\lambda, \mu) \leq \Lambda(\lambda, \mu) + \delta$. Then

$$\begin{aligned} A &\leq \sup_{m \geq M} \mathbf{E}\left[ \mathbf{1}[F_k^2, N_k \geq M] \cdot e^{\sum_{i=2}^{k} \lambda Z_i + \mu E_i} \;\middle|\; N_1 = m \right] \cdot e^{\Lambda(\lambda,\mu)+\delta} \\ &= \sup_{m \geq M} \mathbf{E}\left[ \mathbf{1}[F_{k-1}, N_{k-1} \geq M] \cdot e^{\sum_{i=1}^{k-1} \lambda Z_i + \mu E_i} \;\middle|\; N_0 = m \right] \cdot e^{\Lambda(\lambda,\mu)+\delta}. \end{aligned}$$

An easy induction then shows that

$$\sup_{n \geq M} \mathbf{E}\left[ \mathbf{1}[F_k, N_k \geq M] \cdot e^{\lambda D_k + \mu B_k} \;\middle|\; N_0 = n \right] \leq e^{k\Lambda(\lambda,\mu)+k\delta}.$$

Since $\delta$ was arbitray, the proof is complete. □

The proof of the upper bound of Theorem 5.1 is similar to that of Theorem 4.1 in its structure. Let $\epsilon > 0$. Let $c' = c + \epsilon$, where $c = \sup\{\alpha/\rho : \Lambda^\star(\alpha, \rho) \leq \log d\}$ is the constant defined in the statement of Theorem 5.1. By definition,

$$\mathbf{P}\{H_n > c' \log n\} = \mathbf{P}\{\exists v \in T_n : D_v > c' \log n\}.$$

Recall that $\mathscr{L}_k$ denotes the set of nodes at level $k$ in $T_\infty$. The union bound yields

$$\mathbf{P}\{H_n > c' \log n\} \leq \sum_{k \geq 0} \mathbf{P}\{\exists v \in \mathscr{L}_k : D_v > c' \log n, v \in T_n\}.$$

Using a second union bound over the nodes in each level,

$$\mathbf{P}\{H_n > c' \log n\} \leq \sum_{k \geq 0} d^k \cdot \mathbf{P}\{D_{u_k} > c' \log n, u_k \in T_n\}. \tag{5.2}$$

In order to further bound (5.2), we first restrict our attention to the case $N_{u_k} \geq M$. We have

$$\mathbf{P}\{D_{u_k} > c' \log n, N_{u_k} \geq M\} \leq \mathbf{P}\{(D_{u_k}, B_{u_k}) \in \Gamma, N_{u_k} \geq M\},$$

where $\Gamma = \{(x,y) \in \mathbb{R}^2 : x \geq c'y\}$. By Lemma 5.3, and since $0 \in \mathcal{D}_\Lambda^o$, the upper bound of Gärtner–Ellis theorem (Theorem 2.4) holds: for any $\gamma > 0$, there exists $M_1$ such that for all $M \geq M_1$,

$$\limsup_{k\to\infty} \frac{1}{k} \log \mathbf{P}\left\{D_{u_k} > c' \log n, N_{u_k} \geq M\right\} \leq -\min\left\{\frac{1}{\gamma}, \inf_{(x,y)\in\Gamma} \Lambda^\star(x,y) - \gamma\right\}.$$

By Lemma 4.4, there exists $\beta > 0$ such that $\inf\{\Lambda^\star(x,y) : (x,y) \in \Gamma\} \geq \log d + \beta$. Then, choosing $\gamma < \beta/2$, we have

$$\sum_{k\geq K} d^k \cdot \mathbf{P}\left\{D_{u_k} \geq c' \log n, N_{u_k} \geq M\right\} \leq \sum_{k\geq K} d^k \cdot e^{-k(\beta/2+\log d)} \leq C_1 \cdot e^{-K\beta/2}, \quad (5.3)$$

for all $K \geq K_1$ large enough and some constant $C_1 = C_1(K_1)$,

As in the proof of Theorem 4.1, we treat the values of $k \leq K$ using Markov's inequality. Let $\lambda > 0$, such that $(\lambda, 0) \in \mathcal{D}_\Lambda^o$. There exists $C_2 \geq 0$ and $M_2 > 0$ such that $\sup\{\Lambda^n(\lambda,0) : n \geq M_2\} \leq C_2 < \infty$. Then, for this value of $\lambda$, by Lemma 5.3,

$$\mathbf{P}\left\{D_{u_k} \geq c' \log n, N_{u_k} \geq M_2\right\} \leq e^{kC_2 - \lambda c' \log n}.$$

Therefore, by the union bound,

$$\sum_{k\leq K} \mathbf{P}\left\{\exists v \in \mathscr{L}_k : D_v \geq c' \log n, N_v \geq M_2\right\} \leq \frac{K d^K e^{KC_2}}{n^{\lambda c'}}. \quad (5.4)$$

Let now $M_3 = \max\{M_1, M_2\}$. We have obtained bounds on the terms of (5.3) for every $k$ when $N_{u_k} \geq M_3$. It remains to deal with the nodes at the bottom of the tree for which $N < M_3$. Recall that by assumption, $\mathbf{P}\left\{H_n \geq \psi(n)\right\} = 0$.

$$\begin{aligned}\mathbf{P}\left\{H_n \geq (c+2\epsilon)\log n\right\} &\leq \mathbf{P}\left\{\exists v \in T_n : D_v \geq (c+2\epsilon)\log n - \psi(M_3), N_v \geq M_3\right\} \\ &\leq \mathbf{P}\left\{\exists v \in T_n : D_v \geq (c+\epsilon)\log n, N_v \geq M_3\right\}.\end{aligned}$$

Hence, putting (5.3) and (5.4) together,

$$\mathbf{P}\left\{H_n \geq (c+2\epsilon)\log n\right\} \leq \frac{K d^K e^{KC_2}}{n^{\lambda c'}} + C_1 e^{-K\beta/2}.$$

As $\lambda c' > 0$, this can be made as small as we want by first choosing $K$ and next letting $n$ go to infinity. Since $\epsilon$ was arbitrary, this finishes the proof of the upper bound.

## 5.5 The lower bound

The aim of this section is to build a surviving Galton–Watson process that ensures that nodes with large weighted depth exist in $T_n$ with probability $1 - o(1)$. We split the construction of this process into stages. The proof is similar to that for ideal trees presented in Chapter 4. We rely on Lemma 5.2 and the proof of Theorem 4.1 to show that deep nodes do occur with positive probability, before we boost the probability of their existence to $1 - o(1)$.

SKIMMING THE TREE. Our aim here is to find nodes of sufficiently large weighted depth in $T_n$. Recall that we use the coupled sequence of trees built in section 5.3. We start by finding nodes with large weighted depth in the ideal tree, and then prove that the corresponding nodes in $T_n$ are also sufficiently deep.

**Lemma 5.4.** *Let $T_n$ be a random tree as described in Section 5.2. Let $c = \sup\{\alpha/\rho : \Lambda^\star(\alpha,\rho) \le \log d\}$. For all $\epsilon > 0$, there exists $n_0$ such that*

$$\inf_{n\ge n_0} \mathbf{P}\left\{\exists u \in T_n : D_u \ge (c-\epsilon)\log n\right\} > 0.$$

*Proof.* Let $\epsilon > 0$. By Lemma 4.4, there exists $\alpha$ and $\rho$ such that $\alpha/\rho = c'$ and $I(\alpha,\rho) < \log d$, for some $c'$ such that $c - \epsilon/2 < c' < c$. Let $\alpha$ and $\rho$ be fixed. Let $\ell$ be an arbitrary positive integer to be chosen later. A node $v \in T_\infty$ is called *ideally good* if either it is the root, or $v$ lies $\ell$ levels below an ideally good node $u$ and we have

$$D_v^\infty \ge D_u^\infty + \ell\alpha \quad \text{and} \quad E_v^\infty \le E_u^\infty + \ell\rho.$$

The set of ideally good nodes forms a Galton–Watson tree. Let $Y^\infty$ be the size of the progeny of $u$ in this Galton–Watson process. By linearity of expectation, writing $\pi(u,v)$ for the set of edges on the unique path from $u$ to $v$ in the ideal tree, with $v$ lying $\ell$ levels below $u$,

$$\begin{aligned}
\mathbf{E}Y^\infty &= d^\ell \cdot \mathbf{P}\left\{D_v^\infty - D_u^\infty \ge \alpha\ell, E_v^\infty - E_u^\infty \le \rho\ell\right\} \\
&= d^\ell \cdot \mathbf{P}\left\{\sum_{e\in\pi(u,v)} Z_e^\infty \ge \alpha\ell, \sum_{e\in\pi(u,v)} E_e^\infty \le \rho\ell\right\}.
\end{aligned}$$

By Cramér's theorem (Theorem 2.2), and because of our choice for $\alpha$ and $\rho$, we have

$$\mathbf{E}Y^\infty = d^\ell \cdot e^{-I(\alpha,\rho)\ell+o(\ell)} = e^{\ell\log d-\ell I(\alpha,\rho)+o(\ell)} \xrightarrow[\ell\to\infty]{} \infty.$$

Thus, there exists $\ell$ large enough such that $\mathbf{E}Y^\infty > 1$. This choice makes the process supercritical. Let $\ell$ now be fixed.

Consider now the coupled random trees $T_n$, with size-dependent vectors. A node $v \in T_\infty$ is called *good* if either it is the root, or it lies $\ell$ levels below a good node $u$ and we have

$$E_v \le E_u + \rho\ell \quad \text{and} \quad D_v \ge D_u + \alpha\ell.$$

The set of good nodes is a branching process. However, the progeny distribution $Y_u$ of a node $u$ now depends on $u$ and the process is *not* a Galton–Watson process. We deal with this minor issue using Lemma 3.2. By Lemma 5.2, we have

$$\liminf_{n\to\infty} \mathbf{P}\{Y_u \ge t \mid N_u = n\} \;\ge\; \mathbf{P}\{Y^\infty \ge t\},$$

for all $0 \le t \le d^\ell$. Since $\mathbf{E}Y^\infty > 1$, there exists $M$ large enough that for all $n \ge M$,

$$\mathbf{P}\{Y_u \ge t \mid N_u = n\} \;\ge\; \mathbf{P}\{Y^\infty \ge t\} + \frac{1-\mathbf{E}Y^\infty}{2d^\ell}.$$

Now, by Lemma 3.2, there exists a random variable $Y'$ such that, for all $t$,

$$\mathbf{P}\{Y' \ge t\} = \max\left(\mathbf{P}\{Y^\infty \ge t\} + \frac{1-\mathbf{E}Y^\infty}{2d^\ell}, 0\right).$$

Further, there exist coupled copies of $Y'$, $\{Y'_u, u \in T_\infty\}$ such that have $Y'_u \le Y_u$ if $N_u \ge M$. The Galton–Watson process with progeny distribution $Y'$ is supercritical:

$$\begin{aligned}
\mathbf{E}[Y'] &= \sum_{t=1}^{d^\ell} \mathbf{P}\{Y' \ge t\} \\
&\ge \sum_{t=1}^{d^\ell} \left(\mathbf{P}\{Y^\infty \ge t\} + \frac{1-\mathbf{E}Y^\infty}{2d^\ell}\right) \\
&= \mathbf{E}Y^\infty + \frac{1-\mathbf{E}Y^\infty}{2} \\
&= \frac{1+\mathbf{E}Y^\infty}{2} > 1.
\end{aligned}$$

Therefore, it survives with probability $1 - q > 0$. Note that the garantee $Y'_u \le Y_u$ only occurs if $N_u \ge M$. In particular, it is *not* true that every node $u$ in the coupled Galton–Watson process with progeny distribution $Y'$ is also a good node.

However, in the case of survival, either (a) there is an infinite path of good nodes $u$ with $N_u \ge M$, or (b) there is some good node $w$ with $N_w < M$. Now, if (a) happens, for every integer $k$, there exists a node $v$ such that $D_v \ge \alpha\ell k$ and $N_v \ge M$. So in particular, with

$$k_1 = \left\lfloor \frac{\log n}{\rho\ell} \right\rfloor,$$

$D_v \ge c' \log n - \alpha\ell$, and $v \in T_n$ since $N_v \ge M \ge 1$. In case (b), consider the shallowest good node $w$ such that $N_w < M$. Then, $w$ is part of some generation $k_2$ of the process (at level $k_2\ell$ in $T_\infty$). Since $w$ is good, $M > N_w \ge n\exp(-\rho k_2 \ell)$, and hence,

$$k_2 \ge \frac{\log n - \log M}{\rho\ell}.$$

It follows that $D_w \ge c' \log n - c' \log M$. As a consequence, in both cases, for $n$ large enough, there exists a node $u \in T_n$ with $D_u \ge (c - \epsilon)\log n$, and this happens with probability at least $1 - q > 0$. □

It remains to show that the nodes with large weighted depth found in the previous section do appear in $T_n$ with probability $1 - o(1)$ as $n \to \infty$. Again, we intend to use the standard boosting technique: we run multiple copies of the branching process to increase the chance that one survives. Instead of using the root as a first individual, we want to use some of the $d^t$ nodes at level $t$ as starting individuals of independent processes. However, as for the case of ideal trees, not all such nodes are suitable as starting individuals.

THE NICE PORTION OF THE TREE. Since $\mathbf{P}\{Z = -\infty, E = \infty\}$ may be positive, we cannot expect in general that all $d^t$ nodes at level $t$ are good starting individuals. Indeed, some may not even be fair. In spite of this fact, we claim that under the constraints (i) and (ii), i.e., $\mathbf{P}\{\exists i : Z_i^\infty > -\infty, E_i^\infty < \infty\} = 1$ and $\mathbf{P}\{Z^\infty > -\infty, E^\infty < \infty\} > 1/d$, there are enough of them. In order to prove this claim, we use a second branching process defined on the top $t$ levels.

We look first at the ideal tree. Let $v \in T_\infty$ be called *ideally nice* if either it is the root, or it is linked to an ideally nice node by an edge $e$ and we have

$$Z_e^\infty > a \quad \text{and} \quad E_e^\infty < b.$$

Let $R_t^\infty$ be the number of ideally nice nodes in $\mathscr{L}_t$, the set of nodes $t$ levels away from the root in $T_\infty$. Then $\{R_t^\infty, t \geq 0\}$ is a Galton–Watson process. By hypothesis (ii), $\mathbf{P}\{Z^\infty > -\infty, E^\infty < \infty\} > 1/d$, hence there exist $\delta > 0$, $a_0$ and $b_0$ such that for all $a \leq a_0$ and $b \geq b_0$, $\mathbf{P}\{Z^\infty > a, E^\infty < b\} > 1/d + \delta$. Now, by assumption (i), $\mathbf{P}\{\exists i : Z_i^\infty > -\infty, E_i^\infty < \infty\} = 1$, and thus $\mathbf{P}\{\exists i : Z_i^\infty > a, E_i^\infty < b\} \to 1$, as $a \to -\infty$ and $b \to \infty$. By Theorem 3.5, the process survives with probability at least $1 - q'$, and $q' = q'(a, b)$ can be made as small as we want by choice of $a$ and $b$. If $R_t^\infty > 0$ for all $t \geq 0$, then by Theorem 3.2 $R_t^\infty \to \infty$ as $t \to \infty$ with probability one. As a consequence, for any integer $r$, there exists $t_0$ such that $\mathbf{P}\left\{ R_{t_0}^\infty \leq r \mid R_t^\infty > 0, \forall t \geq 0 \right\} \leq 1/r$.

Let us now go back to the non-ideal random trees, with size-dependent distributions. In the random tree $T_n$, a node at level $t$ is called *nice* if $D_u \geq at$ and $B_u \leq bt$. By Lemma 5.2, the number $R_{t_0}$ of nice nodes $u$ at level $t_0$ satisfies, for $n$ large enough,

$$\mathbf{P}\{ R_{t_0} \leq r \mid R_t^\infty > 0, \forall t \geq 0\} \leq 2/r. \tag{5.5}$$

Observe in particular that the conditioning is meaningful since we consider the coupled sequence of trees. Equation (5.5) gives us the handle we need on the number of nodes we can used as starting individual in the boosting step.

Boosting the survival probability. Let $\epsilon > 0$. Let $\{T_\infty(v_i), 1 \leq i \leq R_{t_0}\}$ be the family of subtrees of $T_\infty$ rooted at the nice nodes $\{v_i, 1 \leq i \leq R_{t_0}\}$. The processes of good nodes described in the proof of Lemma 5.4 evolve independently in every $T_\infty(v_i)$. Furthermore, by Lemma 5.4, there is $n_0$ such that for all $1 \leq i \leq R_{t_0}$ and for all $m \geq n_0$,

$$\mathbf{P}\left\{\exists u \in T_\infty(v_i) : D_u - D_{v_i} \geq \left(c - \frac{\epsilon}{2}\right) \log m, B_u - B_{v_i} < \log m\right\} \geq 1 - q. \tag{5.6}$$

By construction, we have $D_{v_i} \geq at_0$ and $B_{v_i} \leq bt_0$, for $1 \leq i \leq R_{t_0}$. Let $n$ be large enough, and let $m$ be such that $\log m = \log n - bt_0$. If one can find a node $u$ in $T_\infty(v_i)$ as described in (5.6), then

$$D_u \;\geq\; at_0 - \left(c - \frac{\epsilon}{2}\right) t_0 b + \left(c - \frac{\epsilon}{2}\right) \log n \;\geq\; (c - \epsilon) \log n,$$

for $n$ large enough. Such a node $u$ is called a *deep* node. Moreover, $B_u < \log m + B_{v_i} \leq \log n$ so $u \in T_n$ and $H_n \geq D_u \geq (c - \epsilon) \log n$.

If no deep node exists, then one of the following must occur: either $\{R_t, t \geq 0\}$ dies, or it survives but $R_{t_0} \leq r$, or we cannot find a deep node in any of the $R_{t_0} \geq r$ independent trees $T_\infty(v_i)$. As a consequence, for $n$ large enough,

$$\begin{aligned} \mathbf{P}\{H_n \leq (c-\epsilon)\log n\} &\leq \mathbf{P}\{R_{t_0} < r\} + \mathbf{P}\{H_n \leq (c-\epsilon)\log n \mid R_{t_0} \geq r\} \\ &\leq \mathbf{P}\{\exists t \geq 0 : R_t = 0\} + \mathbf{P}\{1 \leq R_{t_0} < r\} + q^r, \end{aligned}$$

by independence of $T_\infty(v_i)$, $1 \leq i \leq R_{t_0}$. It follows that

$$\mathbf{P}\{H_n \leq (c-\epsilon)\log n\} \leq q' + \frac{2}{r} + q^r.$$

This can be made as small as we want by choice of $q' = q'(a, b)$ and $r$. This completes the proof of the lower bound.

## 5.6 The height of trees of effective size $n$

In our model of random tree, we have allowed $Z_e = -\infty$ with positive probability. When this happens for some edge $e$, then we have $D_u = -\infty$ for all $u \in T_\infty$ such that $e \in \pi(u)$. The effect of $Z_e = -\infty$ is to cut a subtree, exactly as $E_e = \infty$ does. In some applications, one is interested in the height of a random tree in terms of its *effective size* $\#T_n$, i.e., the number of nodes that are significant for the height:

$$\#T_n = |\{u \in T_n : D_u > -\infty\}|.$$

The only difference with the height $H_n$ consists in a scaling factor.

**Theorem 5.2.** *Let $T_n$ be a random tree as described in Section 5.2, of (random) effective size $s_n = \#T_n$. Then, its height satisfies $H_n = \frac{c}{\gamma} \log s_n + o(\log s_n)$ in probability, as $n \to \infty$, where $c = \sup\{\alpha/\rho : \Lambda^\star(\alpha, \rho) \le \log d\}$ and $\gamma = -\sup\{\phi : \Lambda_Y(\phi) \le -\log d\}$, with $Y = E + \infty \cdot \mathbf{1}[Z = -\infty]$.*

Theorem 5.2 follows easily from the the following lemma about the effective size of $T_n$. This is a generalization of Lemma 4.9 of the previous chapter.

**Lemma 5.5.** *Let $T_n$ be a random tree as defined in Section 5.2. Let $(Z, E)$ be the limit vector. Then, as $n \to \infty$, $\log \#T_n \sim \gamma \log n$ in probability, where $\gamma = -\sup\{\phi : \Lambda_Y(\phi) \le -\log d\}$.*

*Proof.* The modified size-dependent time random variables are now $\{Y^m, m \ge 0\}$ where $Y^m = E^m + \infty \cdot \mathbf{1}[Z^m = -\infty]$. Upper and lower bounds on $\#T_n$ may be obtained by respectively lower, and upper bounding $Y^n$ so as to have i.i.d. variables, and then using Lemma 4.9. We now describe the upper bound, and omit the proof of the lower bound since it follows the same lines. We have, for all $\phi \in \mathbb{R}$,

$$\begin{aligned}\Lambda_{Y^m}(\phi) &= \log \mathbf{P}\{Z^m > -\infty, E^m < \infty\} + \Lambda^m(0, \phi)\\ &\to \log \mathbf{P}\{Z > -\infty, E < \infty\} + \Lambda(0, \phi),\end{aligned}$$

as $m \to \infty$. Since $0 \in \mathcal{D}^o_\Lambda$, $Y^m \to Y$ in distribution (see Billingsley, 1995, p. 390). We use a coupling argument. Let $F_m$ and $F$ be the distribution functions of $Y^m$ and $Y$, respectively. Let $G_M(x) = \sup\{F_m(x), m \ge M\}$. The function $G_M$ is the distribution function of a proper random variable $W$. By the dominated convergence theorem, we have $\Lambda_W(\gamma + \epsilon) \to \Lambda(\gamma + \epsilon) < \log d$. As a consequence, there exists $M$ large enough that $\Lambda_W(\gamma + \epsilon) \le \log d$.

Now, for $m \ge M$, $Y^m$ stochastically dominates $W_M$. Let $U$ be a $[0, 1]$-uniform random variable. Let $G_M^{-1}$ be the inverse of $G_M$, i.e., the function such that for all $x \in \mathbb{R}$, $G_M^{-1} \circ G_M(x) = x$. By the inverse transform technique (Grimmett and Stirzaker, 2001), for each node $u \in T_\infty$, $F(Y_u)$ is a $[0, 1]$-random variable, and $G_M^{-1} \circ F(Y_u) \le Y_u$ is distributed as $W$. Let $T_n^M$ be the subtree of $T_n$ consisting of nodes $u$ with $N_u \ge M$.

There are at most $\#T_n^M \cdot d$ hanging subtrees with $N_u < M$, each one of effective size at most $M$. It follows that $\#T_n \leq \#T_n^M(1 + dM)$ and

$$\limsup_{n\to\infty} \frac{\log \#T_n}{\log n} \leq \limsup_{n\to\infty} \frac{\log \#T_n^M + \log(1 + dM)}{\log n} \leq \gamma + \epsilon,$$

where the last inequalities follows from the choice of $M$ and Lemma 4.9. Since $\epsilon$ was arbitrary, the proof is complete. □

## 5.7 Applications

We now present some applications of Theorem 5.1. Our goal is to emphasize the wide range of problems that may be handled, even if they apparently are very far from heights of random trees.

### 5.7.1 Variations on binary search trees

Binary search trees probably provide the easiest example of application for Theorem 5.1. Recall that binary search trees (Knuth, 1973c) are search trees built on a set of keys $\{1, 2, \ldots, n\}$. Given a permutation $\{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ of the keys, the first element $\sigma_1$ is stored at the root of a binary tree. The set of keys is then partitioned according to their values into $\{\sigma_i : \sigma_i < \sigma_1\}$ and $\{\sigma_i : \sigma_i > \sigma_1\}$. Both subsets are then treated recursively to form the left and right subtrees of the root, respectively.

If the permutation is taken uniformly at random from the set of permutations of $\{1, \ldots, n\}$, the tree is called a random binary search tree. This model is of great interest, particularly because of its ubiquity in computer science as, e.g., the tree emerging from the branching structure of quicksort (Hoare, 1962). In this model of randomness, $\sigma_1$ is an element of $\{1, \ldots, n\}$ taken uniformly at random and hence the sizes of the left and right subtrees are distributed as $\mathrm{Bin}(n-1, U)$ and $\mathrm{Bin}(n-1, 1-U)$, respectively, where $U$ is a $[0, 1]$-uniform random variable. More precisely, writing $(N_1, N_2)$ for a vector that is distributed as a multinomial$(n - 1; U, 1 - U)$, the vector of interest is

$$\mathcal{X}^n = \left(\left(1, -\log\left(\frac{N_1}{n}\right)\right), \left(1, -\log\left(\frac{N_2}{n}\right)\right)\right). \tag{5.7}$$

One can show that the conditions required to apply Theorem 5.1, are satisfied. In particular:

**Lemma 5.6.** *Let $\mathcal{X}^n$ be defined by (5.7). Let $\mathcal{X} = ((1, -\log U), (1, -\log(1-U)))$. Then, $\Lambda_{\mathcal{X}^n} \to \Lambda_{\mathcal{X}}$ everywhere.*

*Proof.* The weights are irrelevant here, and we consider

$$(E_1^n, E_2^n) = (\log n - \log N_1, \log n - \log N_2)$$

only. Observe that $(N_1, N_2)$ is distributed as $(\lfloor nU \rfloor, \lfloor n(1-U) \rfloor)$, where $U$ is a $[0,1]$-uniform random variable. For all $\mu_1, \mu_2 \in \mathbb{R}$,

$$M_n(\mu_1, \mu_2) \stackrel{\text{def}}{=} \mathbf{E}\left[e^{\mu_1 E_1 + \mu_2 E_2}\right] = \mathbf{E}\left[\left(\frac{\lfloor nU \rfloor}{n}\right)^{-\mu_1} \cdot \left(\frac{\lfloor n(1-U) \rfloor}{n}\right)^{-\mu_2}\right].$$

We have, for all $\mu_1, \mu_2 \in \mathbb{R}$

$$\left(\frac{\lfloor nU \rfloor}{n}\right)^{-\mu_1} \cdot \left(\frac{\lfloor n(1-U) \rfloor}{n}\right)^{-\mu_2} \xrightarrow[n\to\infty]{} U^{-\mu_1} \cdot (1-U)^{-\mu_2}$$

almost surely. Therefore, if $\mu_1 < 1$ and $\mu_2 < 1$, by the bounded convergence theorem,

$$M_n(\mu_1, \mu_2) \to \mathbf{E}\left[U^{-\mu_1} \cdot (1-U)^{-\mu_2}\right].$$

If, on the other hand, either $\mu_1 \geq 1$ or $\mu_2 \geq 1$, then by Fatou's Lemma (see, e.g., Billingsley, 1995),

$$\liminf_{n\to\infty} M_n(\mu_1, \mu_2) \geq \mathbf{E}\left[U^{-\mu_1} \cdot (1-U)^{-\mu_2}\right] = \infty.$$

Thus, we have convergence everywhere in $\mathbb{R} \cup \{+\infty\}$, which completes the proof. ☐

**Remark.** In the following, we will not prove the convergence of the cumulant generating functions any more, and only refer to Lemma 5.6.

Hence, for this model, $E = -\log U$ and $Z = 1$. The random variable $E$ is then distributed as an exponential with mean 1 and Theorem 5.1 immediately implies the following theorem of Devroye (1986).

**Theorem 5.3** (Devroye 1986). *Let $T_n$ be a random binary search tree. Let $H_n$ be its height. Then $H_n \sim c \log n$, in probability as $n \to \infty$, where $c = 1/\rho_o = 4.311\ldots$ and $\rho_o = \inf\{\rho : \rho - 1 - \log \rho \leq \log 2\}$.*

The value $4.311\ldots \log n$ is fairly large compared to $\lfloor \log_2 n \rfloor$, the height of a complete binary tree with $n$ nodes. As this value represents the worst case search time, various methods have been used to shrink it and hence obtain more efficient search trees. Some use splits that are more balanced towards $(1/2, 1/2)$. One way to achieve more balanced splits is to use the median of $2k+1$ keys as a pivot (Van Emden, 1970). When $k$ is fixed, the split at every node is still given by (5.7) but now $(N_1, N_2)$ is distributed as a multinomial $(n-1; U_k, 1-U_k)$ and $U_k$ is a beta$(k+1, k+1)$ random variable. Again, we see that for $\mathcal{X} = ((1, -\log U_k), (1, -\log(1-U_k)))$, $\Lambda_{\mathcal{X}^n} \to \Lambda_{\mathcal{X}}$ everywhere as $n \to \infty$. This suffices for the hypothesis of Theorem 5.1 to hold.

**Theorem 5.4** (Devroye 1993). *Let $T_n$ be a binary search tree built with the medians of $2k+1$ keys as pivots. Then the height $H_n$ of $T_n$ satisfies $H_n \sim c_k \log n$ in probability as $n \to \infty$, where $c_k$ is the unique solution of*

$$\frac{s}{c_k} + \sum_{i=k+1}^{2k+1} \log\left(1 - \frac{s}{i}\right) = \log 2,$$

*and $s$ is implicitly defined by*

$$\frac{1}{c_k} = \sum_{i=k+1}^{2k+1} \frac{1}{i-s}.$$

If $k$ is fixed, we can make $c_k$ close to $1/\log 2$. However, for each $k$ we have $c_k > 1/\log 2$. One can improve this by taking values of $k$ that depend on the number of keys stored in a subtree. If $k \to \infty$ as $n \to \infty$, we see that $\mathcal{X}^n \to \mathcal{X} = ((1, \log 2), (1, \log 2))$ a.s. as $n \to \infty$. Theorem 5.1 then implies that $H_n \sim \log_2 n$, in probability as $n \to \infty$. This strengthens the theorem of Martínez and Roura (2001) which asserts that the average depth, in this case, is asymptotic to $\log_2 n$ (see also Van Emden, 1970).

### 5.7.2 Random recursive trees

The random recursive tree is one of the simplest random trees (Meir and Moon, 1978). One way to describe its construction is by successive insertions of nodes. A recursive tree of size one consists of a single node $v_1$. At each further step $i$ a new vertex $v_i$ is added to the tree and tied to a uniformly random node from $\{v_1, v_2, \dots, v_{i-1}\}$. This is sometimes called a Yule process. Various functionals of this tree have been studied in the literature (Smythe and Mahmoud, 1995). We are particularly interested in its height $H_n$ when $n$ goes to infinity.

**Theorem 5.5** (Devroye 1987, Pittel 1994). *The height $H_n$ of a random recursive tree with $n$ nodes is asymptotic to $e \log n$ in probability as $n$ goes to infinity.*

Random recursive trees have unbounded degree, and it seems that Theorem 5.1 will be of little help. However, one can look at the sequence of *depths* in a random recursive tree as *weighted depths* in a related binary tree (Figure 5.1). Recall that a random binary search tree can be built by choosing, at each step, an external node uniformly at random, and replacing it with an internal one. Therefore, building an auxiliary binary search tree in which the *external nodes* represent the *nodes* of our random recursive tree solves the issue of the uniform choice.

This leads to a construction that maps the nodes of a rooted tree to the external nodes of a binary tree. Consider a rooted tree $\mathcal{T}_n$ on $n$ vertices. Let $\mathcal{M}_2 = \{d_1, d_2, \dots, d_n\}$ be a multiset of numbers that represent the distances from the nodes to the root in $\mathcal{T}_n$. To make the mapping more visual, we also describe the construction of a weighted binary tree $T_n$ on $n$ external vertices together with $M_n$, the sequence of distances in $T_n$ (see Figure 5.1).

- $\mathcal{T}_1$ consists of a single node and $\mathcal{M}_1 = \{0\}$. Appending a node yields a tree on two nodes and $\mathcal{M}_2 = \{0, 1\}$. Let $T_2$ be the binary tree with two external nodes. Let $e$ and $f$ be its edges. Label them with $z_e = 1$ and $z_f = 0$. Consider the labels as distances. Then $T_2$ has distance sequence $M_2 = \{0, 1\} = \mathcal{M}_2$.

- Suppose now that we are given $\mathcal{T}_n$ and the corresponding $T_n$. They match the distance sequence $\mathcal{M}_n = \{d_1, d_2, \ldots, d_n\}$. Appending $v$ to a node $u$ means that we define $\mathcal{M}_{n+1} = \mathcal{M}_n \cup \{d+1\}$, where $d \in \mathcal{M}$ is the distance from $u$ to the root in both $\mathcal{T}_n$ and $T_n$. In terms of trees, we replace the external node $u$ in $T_n$ by an internal node $x$. There are two new external nodes associated with $x$, and the edges $e$ and $f$ out of $x$ are labeled $z_e = 1$ and $z_f = 0$. We may as well label the new external vertices $v$ (such that $e = xv$) and $u$ (with $f = xu$). Then we clearly have $M_{n+1} = M_n \cup \{d+1\}$, and the sequences $\mathcal{M}_{n+1}$ and $M_{n+1}$ match, as required.

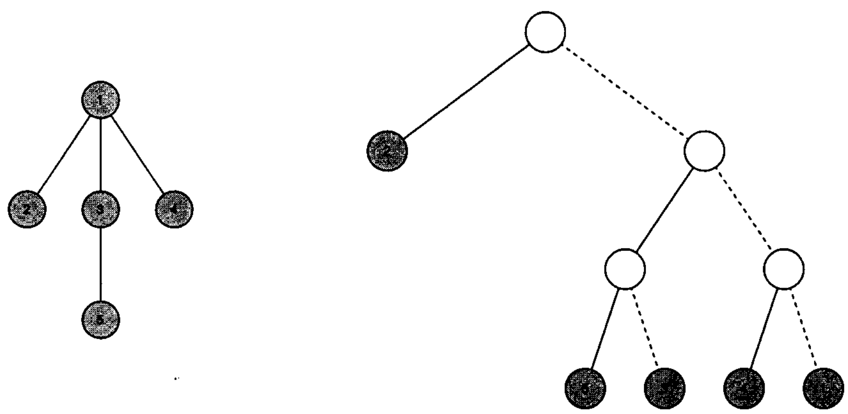

***Figure 5.1:*** *A rooted tree and the corresponding binary tree. The white nodes have been added for the sake of the construction. Solid lines correspond to edges with $Z = 1$ and dashed ones to those with $Z = 0$. Therefore, 1 is equivalent to the root (as the root distance is zero), 2 to the first child of the root (distance one), and so on.*

Replacing deterministic labels by random variables makes this model fit for our framework. For the same reason as in binary search trees, $E =$ exponential(1). Since on any path $\pi$ from the root in $T_\infty$, each edge $e$ is as likely to be labeled with 0 as with 1, we have $Z =$ Bernoulli(1/2).

From Theorem 5.1, we have $H_n \sim c \log n$, where $c = \sup\{\alpha/\rho : \Lambda^\star(\alpha, \rho) \le \log 2\}$. Here (Dembo and Zeitouni, 1998), we have

$$\Lambda^\star(\alpha, \rho) = \alpha \log \alpha + (1-\alpha)\log(1-\alpha) + \log 2 + \rho - 1 - \log \rho,$$

which yields, since the optimum is clearly reached for equality,

$$c = \sup\left\{\frac{\alpha}{\rho} : \alpha \log \alpha + (1-\alpha)\log(1-\alpha) + \rho - \log \rho = 1\right\}. \tag{5.8}$$

The slope $\rho(\alpha)$ of the curve in (5.8) satisfies

$$\frac{d\rho}{d\alpha} = \frac{\log \alpha - \log(1-\alpha)}{1/\rho - 1}. \tag{5.9}$$

Recalling the geometric interpretation shows that the optimal $\alpha$ verifies

$$\frac{d\rho}{d\alpha} \cdot \alpha = \rho.$$

Straightforward manipulations using (5.9) give $\alpha \log \alpha - \alpha \log(1-\alpha) = 1 - \rho$. Taking the value for $1-\rho$ in the equation (5.8) finally gives the desired result, that is, $\alpha/\rho = e$.

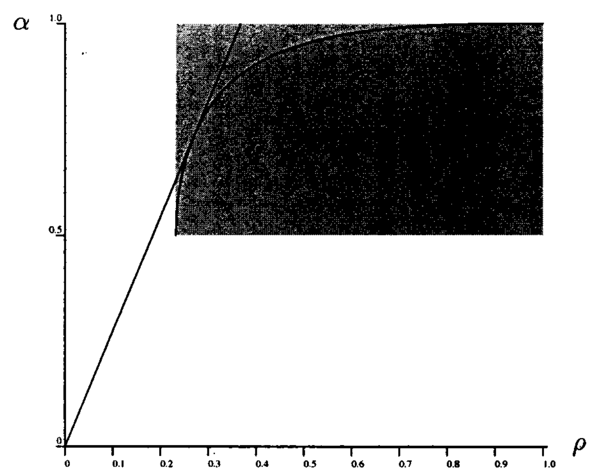


***Figure 5.2:*** *A portion of the level set of interest for random recursive trees* $\Psi(\log 2) = \{(\rho,\alpha) : \alpha \log \alpha + (1-\alpha)\log(1-\alpha) + \rho - \log \rho \le 1\}$.

### 5.7.3 Random lopsided trees

In information theory, researchers are interested in building codes that are optimal with respect to various measures. Prefix-free codes are particularly interesting because they can be decoded directly by following a path in a tree, and output a character corresponding to the codeword when reaching a leaf. In such trees, a node $u$ represents a prefix $p$, and its children the words that can be built by appending a symbol to $p$. In digital applications, characters are usually encoded with bits and therefore, processing each symbol has the same cost. We can think of costs by assigning lengths to the edges in the tree. In this case, they would all have equal lengths. But for some codes the length of codewords is variable. These are called Varn codes (Varn, 1971). Such encodings lead to trees whose edges have non-equal lengths. The corresponding trees are called *lopsided* (see Kapoor and Reingold, 1989; Choi and Golin, 2001).

Let $z_1 \le z_2 \le \cdots \le z_d$ be fixed positive real numbers. Then a tree is said to be lopsided if it is $d$-ary rooted, and for each node, the edge to the $i$-th child has length $z_i$. We now define a model of random lopsided trees, and show that their heights follow from Theorem 5.1. As for random recursive trees, we use a sequential process: start with a tree $T_1$ on single internal node. To build a random tree $T_{n+1}$ with $n+1$ (internal) nodes, take an instance of $T_n$, pick an external node uniformly at random, and replace it with an internal node, exactly as we did in section 5.7.2. The weights of the $d$ child-edges of that internal node are $z_1, z_2, \dots, z_d$. We assume that the $\{z_i, 1 \le i \le d\}$ are not all equal. In this model, $E$ is exponential and $Z = z_I$, where $I$ is uniform on $\{1, \dots, d\}$.

**Theorem 5.6.** *The height $H_n$ of a random $d$-ary lopsided tree with $n$ nodes built with the cost sequence $\{z_1, z_2, \dots, z_d\}$ satifies*

$$H_n = \frac{c}{d-1} \cdot \log n + o(\log n) \quad \textit{in probability,}$$

*as $n \to \infty$, where*

$$c = \sup\left\{\frac{\alpha}{\rho} : \alpha t(\alpha) + \log\alpha - \log\left(\sum_i z_i e^{t z_i}\right) + \rho - 1 - \log\rho \le 0\right\}, \tag{5.10}$$

*and $t(\alpha)$ is uniquely defined by*

$$\sum_{i=1}^{d} (\alpha - z_i) e^{t z_i} = 0. \tag{5.11}$$

**Remark.** Theorem 5.6 does not formally apply to the case of equal $z_i$'s. However, it is easy to verify that when $z_1 = z_2 = \cdots = z_d = 1$, we are led to

$$H_n \sim \frac{c}{d-1} \log n \quad \text{in probability,}$$

as $n \to \infty$, where $c = 1/\rho$, and $\rho$ is the unique solution greater than 1 of $\rho - 1 - \log \rho = \log d$. In particular, for $d = 1$, $c = 4.311\ldots$ since the tree is then a binary search tree.

Our random lopsided trees may also be used when we replace a random node by a fixed deterministic tree. The growing process is as follows. Start with a grey node. Each step sees the replacement of uniformly selected random grey node by a deterministic tree consisting of $k$ nodes (see, e.g., Figure 5.3). In this replacement tree, all leaves, as well as none, some or all of the internal nodes are painted grey (if the root is grey, then the node just replaced may be selected again), for a total of $\ell \leq k$ grey nodes. If we are interested in standard distances to the root, and in the classical definition of the height, then we can imagine another tree in which the replaced node receives a number $\ell$ of children, with edge weights equal to the distances to the root in the replacement tree. The original tree has sizes given by $1 + s(k-1)$ for $s$ integer, and the new imagined tree has sizes given by $1 + s\ell$ for $s$ integer: they are linearly related. The weighted height in the new tree corresponds to the standard height in the original tree. We work out two examples.

**Example.** In Figure 5.3, we replace a randomly picked grey node by a subtree with five nodes, two of which two grey nodes, at distances 1 and 3 from their roots. This corresponds to a random lopsided tree (modulo a proportionality constant in the size of the tree) with weights $(1, 3)$, and fanout $d = 2$. The slope of the tangent going through the origin is $9.3389\ldots$, implying $H_n \sim 9.3389\ldots \log n$ in probability, as $n \to \infty$.

**Example.** In Figure 5.4, we have the same replacement, but paint all five nodes grey. This yields the random lopsided tree with fanout $d = 5$ and cost vector $(0, 1, 1, 2, 3)$. The slope of the optimal tangent is $20.966\ldots$, which gives the height after renormalization: $H_n \sim 5.241\ldots \log n$ in probability, as $n \to \infty$.

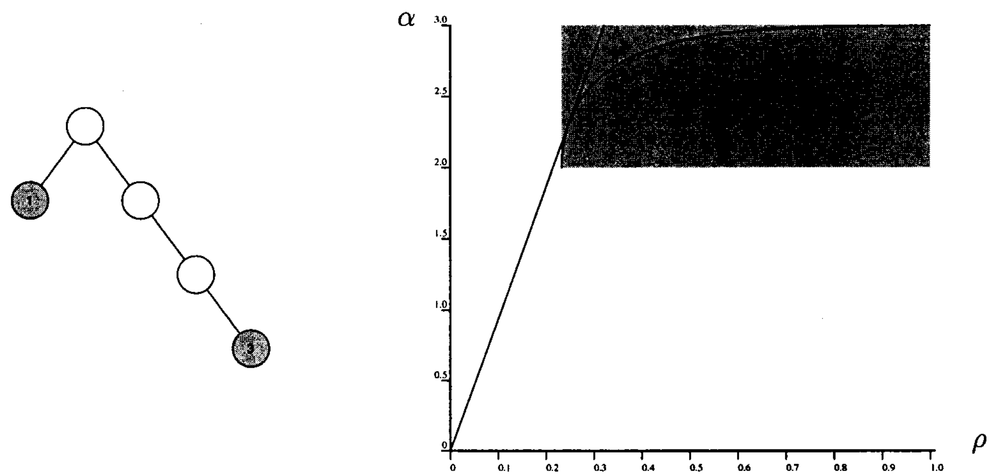


***Figure 5.3:*** *The pattern that replaces a grey node and the portion of interest of* $\Psi(\log 2)$ *together with the optimal tangent when the set of costs is* $\{1, 3\}$. *The nodes are labeled with their depth.*

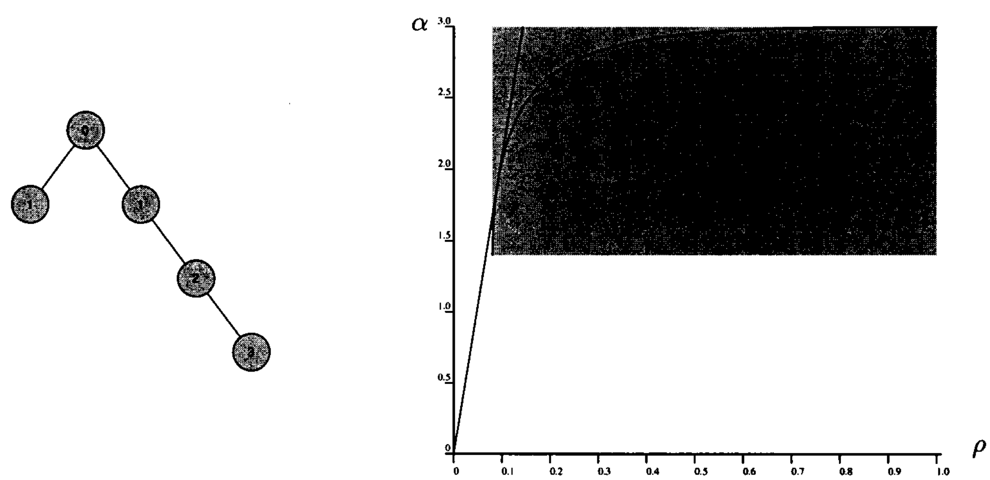


***Figure 5.4:*** *With the set of costs* $\{0, 1, 1, 2, 3\}$, *one can think of a uniform grey node being replaced by the tree pattern on the left.*

*Proof of Theorem 5.6.* In this model, external nodes are picked uniformly at random and $E$ is expoential. Since on a path to the root, each edge is equally likely to have any cost in $\{z_1, \ldots, z_d\}$, and by independence of $Z$ and $E$,

$$\Lambda(\lambda, \mu) = \log \mathbf{E}\left[e^{\lambda Z + \mu E}\right] = \log \sum_i e^{t z_i} - \log d + \log \mathbf{E}\left[e^{\mu E}\right].$$

Using the definition for $\Lambda^\star$, we see that we see that the optimal value is obtained for

$$\alpha = \frac{\sum_i c_i e^{tz_i}}{\sum_i e^{tz_i}},$$

which is equivalent to (5.11). The value $t(\alpha)$ is unique as long as at least two of the $z_i$'s are distinct. The constraint (5.10) follows immediately from Theorem 5.1. □

### 5.7.4 Plane oriented, linear recursive and scale-free trees

Plane oriented trees (or plane recursive trees or PORTs) are an ordered version of recursive trees due to Szymański (1987). They may be obtained using successive insertions as well. The difference lies in that a parent is no longer chosen uniformly, but rather with probability proportional to one plus its outdegree. This is also the preferential attachment model used by Barabási and Albert (1999) to represent the web, and a particular case of the more general recursive trees of Pittel (1994).

Plane oriented trees (PORTs) are rooted trees in which the children of every node are oriented. A random PORT with $n$ nodes is defined as a tree taken uniformly at random from the set of $(n-1)!$ plane oriented trees with $n$ nodes. The depths of nodes in random PORTs have been studied by Mahmoud (1992a) and the height by Pittel (1994). An interesting property of PORTs is their recursive description: one can view a random PORT with $n$ nodes as a random PORT with $n-1$ nodes, to which we add a node uniformly at random in the set of slots available. Nodes have labels 1 through $n$ in order of addition, and therefore, the labels are always increasing on paths down from the root. The slots are the positions in the tree that lead to different new trees. Because of the order, each node with $k$ children has $k+1$ slots (external nodes) attached to it as described in Figure 5.5.

We may consider them as *linear recursive trees*, a more general model of Pittel (1994), which has also been dealt with by Biggins and Grey (1997). For this kind of tree, each node $u$ has a weight $w_u$, and when growing a random linear recursive tree, a new node is added as a child to a node $u$ picked at random with probability

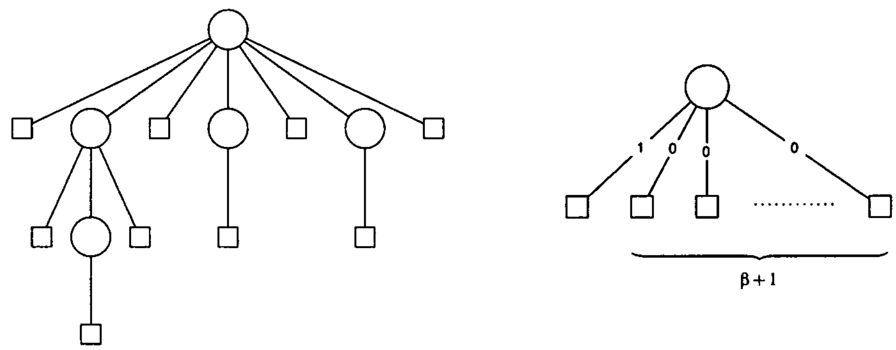


*Figure 5.5: A PORT with the slots represented by squares on the left and the tree pattern on the right, representing the replacement of an external node. The labels on the edges are the costs of crossing them.*

proportional to $w_u$. For linear recursive trees, we have $w_u = 1 + \beta \mathrm{deg}_u$, where $\mathrm{deg}_u$ denotes the number of children of $u$ and $\beta \geq 0$ is called the parameter. We can obtain the same distribution on trees by taking external nodes uniformly at random and with a suitable number of external nodes for each vertex, at least when $\beta$ is integer (see below).

Assume that $\beta$ is integer-valued. It is easily seen that when we pick a uniform external node at depth $d$, and replace it by $\beta + 2$ new external nodes, $\beta + 1$ at depth $d$ and one at $d + 1$, then this may be seen as replacing a uniform external node by the fixed tree pattern of Figure 5.5. The values of $Z$ for the $\beta + 2$ child-edges of a node consist of one 1 and $(\beta + 1)$ 0's. Therefore, a typical $Z$ is distributed like Bernoulli$(1/(\beta+2))$. One may apply our result on random lopsided trees with fanout $\beta + 2$ to find a new proof of Pittel's theorem on the height of linear recursive trees.

**Theorem 5.7** (Pittel 1994). *Assume that $\beta$ is integer-valued. The height $H_n$ of a random linear recursive tree with parameter $\beta$ and $n$ nodes is such that*

$$\frac{H_n}{\log n} \xrightarrow[n\to\infty]{} \frac{c}{\beta+1}$$

*in probability, as $n \to \infty$ where*

$$c = \sup\left\{\frac{\alpha}{\rho} : \alpha \log \alpha + (1-\alpha)\log\left(\frac{1-\alpha}{\beta+1}\right) + \rho - 1 - \log \rho = 0\right\}.$$

The special case of random recursive trees is obtained for $\beta = 0$ and plane oriented trees for $\beta = 1$ yielding an asymptotic height of $1.7956\ldots \log n$.

### 5.7.5 Intersection of random trees

We can also apply Theorem 5.1 to the intersection of random trees. One can take $k$ independent copies of a certain kind of random $d$-ary tree on $n$ nodes and ask about the height of the intersection (a node is in the intersection if it is present in all $k$ trees). This model was treated by Baeza-Yates et al. (1992) for random binary search trees in the context of tree matching properties arising in the tree shuffle algorithm (Choppy et al., 1989). The authors were in particular interested in the size of the intersection of two random binary search trees. We will consider the intersection of $k$ binary search trees, and of $k$ plane oriented trees.

Let $C_{k,n}$ be a collection of $k$ independent copies of identically distributed random trees with $n$ nodes, and let $T_{k,n}$ be their intersection. Recall that the shape of the random tree in our framework is related to the random variables $E_e$ in all $k$ copies. The random variables $E$ of Theorem 5.1 are now $k$-vectors of independent random variables. From now on, we write $E$ for a coordinate of this vector, and this corresponds to the random variable describing one of the random trees. By independence of the $k$ trees in $C_{k,n}$ the rate function that corresponds to the presence of a node in $T_{n,k}$ is $k\Lambda_E^\star$. We obtain that the rate function to be considered is $\Lambda^\star(\alpha, \rho) = \Lambda_Z^\star(\alpha) + k \cdot \Lambda_E^\star(\rho)$, where $E$ and $Z$ are the random variables describing one single random tree. As an example, this yields the following result.

**Proposition 5.1.** *The height $H_{k,n}$ of the intersection $T_{k,n}$ of $k$ independent copies of*
*(a) random binary search trees is asymptotically $c_{BST}(k) \log n$, in probability, where*

$$c_{BST}(k) = \sup\left\{\frac{1}{\rho} : \rho - 1 - \log\rho \le \frac{\log 2}{k}\right\};$$

*(b) plane oriented trees is asymptotically $(c_{PORT}(k)/2) \log n$, where*

$$c_{PORT}(k) = \sup\left\{\frac{\alpha}{\rho} : \alpha\log\alpha + (1-\alpha)\log\left(\frac{1-\alpha}{2}\right) + k(\rho - 1 - \log\rho) \le 0\right\}.$$

**Remark:** Note that $T_{k,n}$ is likely to contain fewer than $n$ nodes, and the height is not given as a function of the size of $T_{k,n}$.

Table 5.1 gives numerical values of $c_1$ and $c_2$ for certain values of $k$. The limit values as $k \to \infty$ can also be derived.

| | $k$ | | | | |
|---|---|---|---|---|---|
| | 2 | 5 | 10 | 50 | 100 |
| $c_{\mathrm{BST}}$ | 2.62729... | 1.78088... | 1.48726... | 1.18680... | 1.12760... |
| $c_{\mathrm{PORT}}$ | 2.03950... | 1.39752... | 1.20841... | 1.05078... | 1.02788... |

***Table 5.1:*** *Some numerical values of the asymptotic height of $T_{k,n}$.*

**Proposition 5.2.** *There exist limits of both constants $c_{BST}(k)$ and $c_{PORT}(k)$ as $k$ goes to infinity and*

$$\lim_{k\to\infty} c_{BST}(k) = \lim_{k\to\infty} c_{PORT}(k) = 1.$$

**Remark.** Observe in particular that the height of the intersection *does not* converge to the fill up level for binary search trees, which may appear surprising at first glance.

*Proof of Proposition 5.2.* For random binary search trees, this is easily seen since $\{\Lambda^\star(1,\rho) = \rho - 1 - \log\rho = \log 2/k\}$ is the intersection of two explicitly defined curves. By continuity of $\Lambda^\star(1,\rho)$ on $\mathcal{D}^\circ_{\Lambda^\star}$, $\rho \to 1$ as $k \to \infty$.

Consider now PORTs. From the geometric properties of $\{\Lambda^\star(\alpha,\rho) \le \log 3\}$, $\rho \ge \rho_{\min}$, where $\rho_{\min}$ is the value at $\alpha = \mathbf{E}Z = 1/3$, and

$$\rho_{\min} - 1 - \log\rho_{\min} = \frac{\log 3}{k},$$

giving that $\rho_{\min} \to 1$ as $k \to \infty$. As a consequence, we need only look at $\alpha$. Now, the line $\Delta$ going through the origin and $(\rho,\alpha) = (1,1)$ crosses $\{\Lambda^\star(\alpha,\rho) \le \log 3/k\}$ because of its convexity and horizontal tangent at $\rho = 1$. Therefore, the slope of the tangent $\tau$ at the optimal point $(\rho,\alpha)$ is greater than 1. Writing $(\rho_{\min},\alpha_\Delta)$ for the intersection of $\Delta$ and $\{\rho = \rho_{\min}\}$ (Figure 5.6), we get that $\alpha \geq \alpha_\Delta = \rho_{\min}$, yielding $c_{\mathrm{PORT}} \to 1$ as $k \to \infty$. $\square$

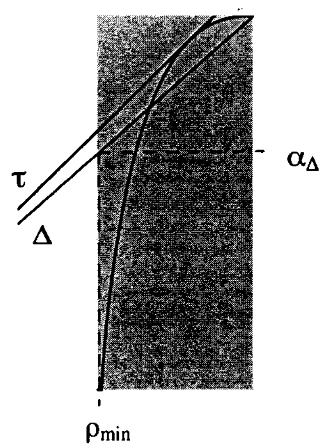


*Figure 5.6:* *$\Psi(\log 3/k)$ together with the optimal tangent $\tau$ and the line $\Delta$ through the origin and $(1, 1)$.*

### 5.7.6 Change of direction in random binary search trees

Consider $T_n$ a rooted binary tree on $n$ nodes. For a node $u \in T_n$, let $D_u$ be the number of changes of direction in on $\pi(u)$, the unique path from the root down to the $u$. If we let 0 and 1 encode a move down to the left and to the right, respectively, then the node $u$ whose path is encoded by 0100101 will have $D_u = 5$, that is, a count of each occurrence of the patterns 01 and 10. We are interested in the maximal value over all the paths of the tree $H_n = \max\{D_u : u \in T_n\}$. When $T_n$ is a random binary search tree, this turns into a random variable that may be handled by our framework. It suffices to notice that if we take a left step, the next move will increase $D$ only if we go right. We have of course something similar when the first step was to the right. Thus, we label the edges as follows. For each level $k \geq 2$ of edges, we form the word $(0110)^{k-1}$, and map the binary characters to the edges from left to right. Then, for a node $u$, $D_u$ corresponds exactly to the sum of these labels along $\pi(u)$ (Figure 5.7).

This means that for the tree to match our model we need $Z$ to be Bernoulli$(1/2)$, and $E$ exponential$(1)$, because the underlying tree is a binary search tree. Therefore

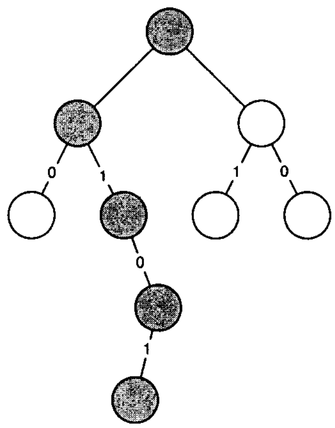


*Figure 5.7: The path consisting of grey nodes is the one with the maximum number of change of direction. Note that the number of changes of direction is the sum of the labels along the path.*

the maximum number of changes of directions along a path in a random binary search tree is asymptotic to the height of random recursive trees.

**Proposition 5.3.** *The maximal number $H_n$ of change of direction along a path in a random binary search tree is asymptotic to* $\mathrm{e} \log n$ *in probability, as* $n \to \infty$.

### 5.7.7 Elements with two lifetimes

Consider a binary tree in which elements have two independent exponential(1) lifetimes, $Z$ and $E$, and let $D_u$ and $B_u$ keep their meaning. In the tree $T_n$, that is, the tree of all nodes $u$ with $B_u \le n$, it is interesting to ask about the maximal value of $D_u$ when measured with respect to the second lifetimes ($Z$). Since $Z$ and $E$ have similar Cramér functions, and both have mean one, we have by Theorem 5.1,

**Proposition 5.4.** *The maximal age $D_u$ of any node $u$ in the tree of two lifetimes described above, cut off at date of birth $B_u \le n$ is $H_n$. We have*

$$\frac{H_n}{n} \xrightarrow[n\to\infty]{} c = 5.82840157\ldots$$

*in probability, where the constant $c$ is defined by*

$$c = \sup\left\{\frac{\alpha}{\rho} : \rho - 1 - \log \rho + \alpha - 1 - \log \alpha \le \log 2\right\}.$$

Thus, in spite of the fact that, when measured by first lifetimes, all have age less than $n$, there exist elements whose age, when measured in the other time scale, is almost six times as large!

### 5.7.8 Random $k$-coloring of the edges in a random tree

Assume that we randomly color the edges of a random binary search tree of size $n$ with $k$ colors, and that we ask for the maximal number $H_n$ of similar colors on one path from a root to a leaf. This is equivalent, when $k$ is constant, to studying the maximum number of red colored edges on such paths. But then, this can be studied by attaching to edges independent copies of $Z$ where $Z = 1$ with probability $1/k$ and $Z = 0$ otherwise. That is, $Z$ is Bernoulli$(1/k)$. We have seen already the rate functions for Bernoulli and exponential random variables (Dembo and Zeitouni, 1998). Then, $\Lambda^\star(\alpha,\rho) = \Lambda^\star_Z(\alpha) + \Lambda^\star_E(\rho)$, where $\Lambda^\star_E(\rho) = \rho - 1 - \log\rho$ and

$$\Lambda^\star_Z(\alpha) = \alpha\log(\alpha) + (1-\alpha)\log\left(\frac{1-\alpha}{k-1}\right) + \log k,$$

and we have $H_n \sim c\log n$, where

$$c = \sup\left\{\frac{\alpha}{\rho} : \alpha\log(\alpha) + (1-\alpha)\log\left(\frac{1-\alpha}{k-1}\right) + \rho - 1 - \log\rho \le \log\left(\frac{2}{k}\right)\right\},$$

Note that for $k = 2$, or $p = 1/2$, we have a situation not unlike that of the maximum number of sign changes in random binary search trees, or the random recursive tree, where the asymptotic maximum value is $e\log n$. Also, clearly, the maximal number of identical colors on a path decreases with the number of colors.

For $k = 1$ and 2 we have the known results for the height of the random binary search trees and random recursive trees, respectively, as one can check in Table 5.2. Clearly, we may even introduce $p$ values not equal to $1/k$, and ask on which path we have most red-blue color changes, for example, where red and blue occur with probabilities $p$ and $q$ respectively.

| | $k$ | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| $c_k$ | 4.3110... | 2.7182... | 2.1206... | 1.7955... | 1.5869... |
| | $k$ | | | | |
| | 6 | 7 | 8 | 9 | 10 |
| $c_k$ | 1.4397... | 1.3292... | 1.2426... | 1.1725... | 1.1148... |

*Table 5.2: Some numerical values of $c_k$.*

**Remark.** To study the maximal number of colors of one kind (among $k$ colors) in a *random recursive tree* instead, it takes just a moment to see that it suffices to take $Z = \text{Bernoulli}(1/k)\times$ Bernoulli$(1/2)$. In other words, $Z$ is Bernoulli$(1/(2k))$.

### 5.7.9 The maximum left minus right exceedance

Let the *differential depth* of a node $u$ be

$$D_u = \sum_{e\in\pi(u)} L(e) - R(e),$$

where $L(e)$ is the indicator of $e$ being a left edge and $R(e)$ is the indicator of $e$ being a right edge. We want to study the extreme value (differential height) $H_n$ of $D_u$, when $u$ ranges over the nodes of a random binary search tree of size $n$. We have seen that for a random binary search tree, $E$ is an exponential random variable with mean one, so $\Lambda_E^\star(\rho) = \rho - 1 - \log\rho$. For this purpose, we may make $Z = 1$ or $-1$ with probability $1/2$. Note that for our $Z$,

$$\Lambda_Z(\lambda) = \log\left(e^\lambda + e^{-\lambda}\right) - \log 2.$$

And we obtain the Crámer function associated to $Z$,

$$\Lambda_Z^\star(\alpha) = \begin{cases} \infty & \alpha \ge 1 \\ \frac{\alpha}{2}\log\left(\frac{1+\alpha}{1-\alpha}\right) + \log 2 - \log\left(\sqrt{\frac{1+\alpha}{1-\alpha}} + \sqrt{\frac{1-\alpha}{1+\alpha}}\right) & 0 \le \alpha < 1. \end{cases}$$

Then, $\Lambda^\star(\alpha,\rho) = \Lambda_Z^\star(\alpha) + \rho - 1 - \log\rho$. Theorem 5.1 allows to conclude that there exists a constant $c$ such that $H_n \sim c\log n$ in probability as $n$ tends to infinity. Numerical tools allow to determine $c = 2.07345\dots$.

### 5.7.10 Digital search trees

This example is similar to the one of Broutin et al. (2006). We consider tries on a finite alphabet $\mathcal{A} = \{1, 2, \dots, d\}$ with the Bernoulli model of randomness: each datum consists of an infinite sequence $A^i = A_1^i, A_2^i \dots$ of i.i.d. random elements of $\mathcal{A}$ (Fredkin, 1960; Szpankowski, 2001). A string $A^i$ corresponds to an infinite path in a $d$-ary tree defined in the following way: from the root, take the $A_1^i$-th first child, next the $A_2^i$-th, and so forth. We prune the subtrees of each node that contain only one single string. The remaining tree is the trie associated with the $n$ strings.

For tries, there is no deterministic bound on the height of a trie built from $n$ or even two strings: Neither Theorem 4.1 nor Theorem 5.1 applies to tries. Various techniques have been used to shrink the height of tries such as PATRICIA (Morrison, 1968) and digital search trees (Coffman and Eve, 1970; Konheim and Newman, 1973). See also the recent survey by Flajolet (2006). We now focus on digital search trees. First, the term of *digital search tree* seems misleading to us, since digital search trees are not *search trees*, where a search query is carried over using the values stored in the nodes. We prefer the term *pebbled tries*, to emphasize the trie structure: a string (a "pebble") is assigned to each node in the tree instead of to each leaf. In this "pebbled" version of tries, a string, taken at random, is associated to the root. Then, the $n-1$ remaining strings are distributed to the $k$ subtrees depending on the value of their first character. The tree is then built recursively.

In a computer, the characters are coded in binary. The *cost* of a character in terms of bit comparisons is then the length of its binary code. The model of pebbled tries has been studied by Broutin et al. (2006) in the case where all $k$ characters have the same cost. However, if one uses an optimal code (one that minimizes the costs of the characters), the lengths of the codewords depend on the character, and hence the costs of characters vary. Also, in such a code, the length of a codeword is obviously dependent of the probability that the corresponding character occurs (prefix codes of Huffman, 1952). Hence, this model of pebbled tries built with Huffman coded characters is a perfect application for Theorem 5.1. Compare with the lopsided trees of section 5.7.3.

Let $p_i$ be the probability that character $i$ occurs at some fixed position of a string. Let $\ell_i$ be the length of the binary codeword for character $i$. Then, at a node $u$ with $N_u = n+1$, the split $\mathcal{V}^n$ is distributed as a multinomial$(n, p_1, p_2, \ldots, p_d)$ random vector. The weights $(Z_1, \ldots, Z_d)$ are deterministic and equal to $(\ell_1, \ell_2, \ldots, \ell_d)$. Now, $\mathcal{V}^n \to (p_1, p_2, \ldots, p_d)$ almost surely, and hence it is easily checked that the required conditions on the random variables are satisfied with $X = (\ell_K, p_K)$ where $K$ is uniform in $\{1, \ldots, d\}$. It follows that

$$\Lambda(\lambda, \mu) = -\log k + \log\left(\sum_{i=1}^{d} e^{\lambda \ell_i - \mu \log p_i}\right).$$

Also, since for all $i$, $\ell_i > 0$ and $\log p_i < 0$ (or there is a.s. only one character in the alphabet and the tree is degenerate), $e^{\Lambda + \log d}$ is a sum of positive convex functions whose gradient spans $(0, \infty)^2$. As a result, for $\alpha, \rho \in (0, \infty)$, there exist $\lambda$ and $\mu$ for which $\sup_{\lambda', \mu'}\{\lambda'\alpha + \mu'\rho - \Lambda(\lambda', \mu')\} = \lambda\alpha + \mu\rho - \Lambda(\lambda, \mu)$ which are given implicitly by

$$\alpha = \frac{\sum_{i=1}^{d} \ell_i e^{\lambda \ell_i} p_i^{-\mu}}{\sum_{i=1}^{d} e^{\lambda \ell_i} p_i^{-\mu}} \qquad \text{and} \qquad \rho = \frac{-\sum_{i=1}^{d} \log p_i e^{\lambda \ell_i} p_i^{-\mu}}{\sum_{i=1}^{d} e^{\lambda \ell_i} p_i^{-\mu}}.$$

Then, by Theorem 4.1, the height of the pebbled trie is asymptotic to $c \log n$ in probability, where $c$ the maximum value of $\alpha/\rho$ along the curve

$$\lambda\alpha + \mu\rho = \log\left(\sum_{i=1}^{d} e^{\lambda \ell_i - \mu \log p_i}\right).$$

Numerical values can easily be obtained for every set of parameters $\{(p_i, \ell_i), 0 \le i \le d\}$.

### 5.7.11 Pebbled TST

In the same vein, we can study the height of a pebbled version of ternary search tries (TST). The (non-pebbled) TST structure introduced by Bentley and Sedgewick (1997) uses early ideas of Clampett (1964) to improve on array-based implementations of tries. If an array is used to implement the branching structure of a node, the number of null pointers can become an issue when the alphabet is large. In TSTs, instead of the usual array, the node structure consist of a binary search tree (BST), therefore forcing small branching factors and limiting the amount of null pointers. So the TST is a hybrid structure combining tries and binary search trees. The high level structure is still that of a trie. Only the structure of a node and the way character matching are handled changes. TSTs have been studied by Clément, Flajolet, and Vallée (1998, 2001).

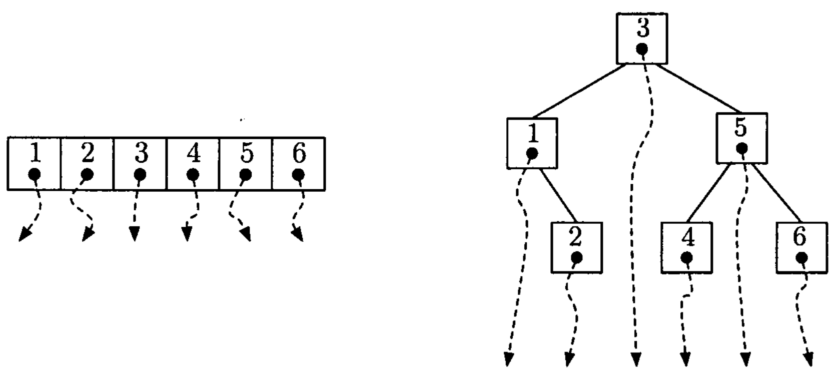


***Figure 5.8:*** *The structure of a node in an array-based trie (left) and TST (right) over the alphabet $\{a, b, c, d, e, f\}$. The pointers used for the high level trie structure are dashed.*

We now describe the modified pebbled version. Let $\{A^1, A^2, \ldots, A^n\}$ be the set of strings, with $A^i = A^i_1 A^i_2 \ldots$, for all $i$. We distinguish the *nodes* of the trie structure from the *slots* of the local binary search trees. As shown in Figure 5.8, each node contains $k$ slots. The nodes at distance the same $j$ from the root are said to be *at level* $j$. At level $j$, the key used for the comparisons is the $j$-th character of the sequences.

The tree is built by assigning the sequences to the first empty *slot* as they come along in order $A^1, A^2, \ldots$. The first string $A^1$ is stored in the first slot of the root of the TST, and partitions the following sequences with respect to their first characters $a$, whether $a < A_1^1$, $a = A_1^1$ or $a > A_1^1$. Given the TST built from the first $m-1$ sequences, $A^m$ moves down the tree as indicated by the sequences stored in the slots encountered, the comparisons being done on the $j$-th character at level $j$. It changes level only upon finding a matching character, in the other cases, it moves in the slots of the *same* node until it eventually finds either an empty slot, or a matching character.

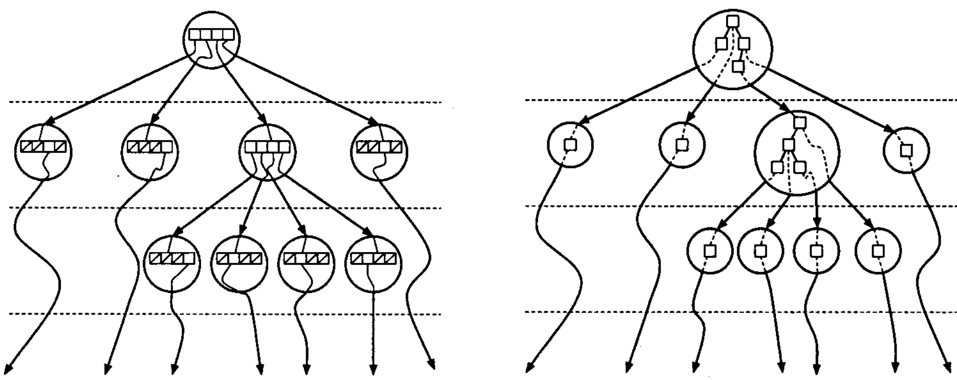

***Figure 5.9:*** *The outer structure of trie (left) and the expanded binary search tree structure of the nodes of a TST on an alphabet of size four. The* nodes *are shown as circles whereas the* slots *are represented by squares.*

We now assume that the strings are independent sequences of i.i.d. characters where a character $a \in \{1, \dots, d\}$ has probability $p_a > 0$. We are interested in the height of a pebbled TST built from $n$ of these independent sequences. Consider a node $u$ whose subtree stores $n+1$ strings. As in the previous section, the split vector at $u$, $(N_1, N_2, \dots, N_d)$, is clearly multinomial$(n, p_1, \dots, p_d)$. Looking at the high level trie structure, the edges may be seen as being weighted by the number of edges in the local binary search tree structure (Figure 6.11). Clearly, the cost of the edge leading to a character $a$ is the 1 plus the depth of the node labeled $a$ in the BST of the node considered. Let $Z_a^n$ be the random variable accounting for this value. Then the vector of interest is

$$\mathcal{X}^n = \left(\left(Z_1^n, -\log\left(\frac{N_1}{n}\right)\right), \dots, \left(Z_d^n, -\log\left(\frac{N_d}{n}\right)\right)\right).$$

The random variable $Z_a^n$ has been studied by Clément et al. (1998) and Archibald and Clément (2006). In particular they studied the expected values and variances of $\{Z_a^n, 1 \leq a \leq d\}$. However, we need information about the distributions of $Z_a^n$ and their limits as $n \to \infty$. Let $\tau$ be the smallest $n$ for which $\{A_1^i, 0 \leq i \leq n\}$ contains a copy of each character. Then, for each $n \geq \tau$, the distribution of $Z_a^n$ is that of $Z_a = Z_a^\tau$, independent of $n$. The random variable $\tau$ is a stopping time and $\mathbf{P}\{\tau \geq n\} \leq (1-\min\{p_i, 1 \leq i \leq d\})^n$. This proves that $\tau$ is a.s. finite and that $Z_a^n \to Z_a$, in distribution. Then, with $\mathcal{X}$ distributed as $((Z_1, -\log p_1), \ldots, (Z_d, -\log p_d))$, one can show that $\Lambda_{\mathcal{X}^n} \to \Lambda_{\mathcal{X}}$ everywhere as $n \to \infty$.

This is sufficient for Theorem 5.1 to apply. The height of the pebbled TST is asymptotic to $c \log n$ in probability, where $c$ is the the maximum value of $\alpha/\rho$ in $\{(\alpha, \rho) : \Lambda^\star(\alpha, \rho) \leq \log d\}$, and $\Lambda^\star$ is the Cramér function associated with $X = (Z_K, -\log p_K)$ where $K$ is uniform in $\{1, \ldots, d\}$. Numerical values can be obtained for examples of $\{p_i, 1 \leq i \leq d\}$.

**Remark.** The height of the non-pebbled version of TST requires more care and is treated in Chapter 6.

### 5.7.12 Skinny cells in $k$-d trees

We consider the $k$-d tree introduced by Bentley (1975). This geometric structure generalizes binary search trees to multidimensional data sets. Given a set $\mathcal{D}$ of $d$-dimensional data points $\{Y^1, Y^2, \ldots, Y^n\}$, where $Y^i = (y_1^i, \ldots, y_d^i)$ for all $i$, we recursively build the following binary tree structure partitioning the data set using comparisons of some of their components. The first datum $Y^1$ is stored at the root. The remaining of the data are processed as follows: $\{Y^i : i \geq 2, y_1^i \leq y_1^1\}$ and $\{Y^i : y_1^i > y_1^1\}$ are assigned respectively to the left and right subtrees, and both subtrees are recursively built using the same method. The comparisons are done in a cyclical way depending on the depth of the node at which they occur: the key used at a node at depth $\ell$ is the $(\ell \bmod d + 1)$-st component of a vector. For a more complete account on $k$-d trees see Gonnet and Baeza-Yates (1991) or Samet (1990a,b).

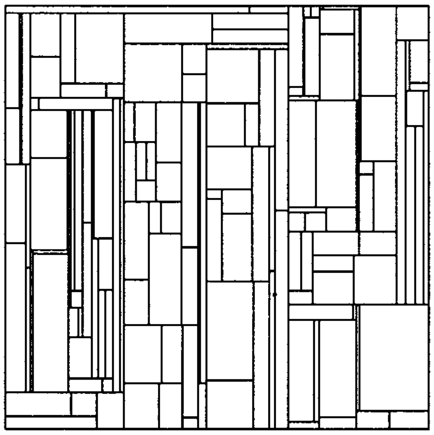

***Figure 5.10:*** *A (randomly generated) $k$-d tree on 150 points uniformly distributed in* $[0,1]^2$.

When the data points are i.i.d. $[0,1]^d$-uniform random variables, one can see a $k$-d tree as a random refining partition of $[0,1]^d$. The root represents $[0,1]^d$, and more generally, a node $u$ represents the set of points $x \in [0,1]^d$ that would be stored in its subtree if they were data points inserted after $u$. Therefore, each cell is split into two along a dividing line, on which lies one of the points $Y^i$, and whose direction changes in a cyclical way. The cells are obviously rectangular. Let $C_u$ be the cell associated with a node $u$. Let $L_1(u), L_2(u), \ldots, L_d(u)$ be the its lengths with respect to the $d$ dimensions. We are interested in the worst case ratio of two dimensions of a cell. For example, if $d = 2$, this is the worst case ratio length over width. By symmetry, since $d$ is bounded, we can always consider the worst case of the first two dimensions, $L_1$ and $L_2$. Such a parameter is of great importance in applications. Indeed, for partial match queries, the running times of algorithms depend on the shape of the cells, and in particular on how close they are to squares (Flajolet and Puech, 1986; Martínez et al., 2001; Devroye et al., 2001). We prove the following:

**Theorem 5.8.** *Let $T_n$ be a $k$-d tree built from $n$ i.i.d. $[0,1]^d$-uniform random points. Let*

$$R_n \stackrel{def}{=} \max\left\{\frac{L_1(u)}{L_2(u)} : u \in T_n\right\}.$$

*Then $R_n = n^{c_d+o(1)}$ in probability, as $n \to \infty$. Furthermore, $c_d$ is the maximum value of $\alpha/\rho$ in the set $\{(\alpha,\rho) : \lambda\alpha + \mu\rho + \log((1-\mu)^2 - \lambda^2) + (d-2)\log(1-\mu) \le d\log 2\}$, where*

$$\mu = 1 - \frac{\rho(d-1) + \sqrt{\rho^2 + \alpha^2 d(d-2)}}{\rho^2 - \alpha^2} \quad \text{and} \quad \lambda = -\alpha \cdot \frac{d(d-2) - 2\rho(d-1)(1-\mu)}{(\rho^2-\alpha^2)(\rho(1-\mu) - (d-2))}.$$

*In particular, $c_d < 1$ for all $d \ge 2$.*

| | $d$ | | | | | |
|---|---|---|---|---|---|---|
| | 2 | 3 | 5 | 10 | 40 | 100 |
| $c_d$ | 0.86602... | 0.79047... | 0.71246... | 0.63483... | 0.54976... | 0.52442... |

***Table 5.3:*** *Some numerical values for the constant $c_d$ describing the asymptotic values of $R_n$ in a d-dimensional k-d tree.*

*Proof.* We intend to express the maximum ratio as a weighted height. Since in $k$-d trees, *not* all the levels in the tree are equivalent, we proceed in two stages: we first only consider the levels with depth $0 \bmod d$; next we only need to consider the levels $1 \bmod d$ since for the other $d-2$ levels, the ratio $L_1/L_2$ is not modified.

If we group the levels by bunches of $d$, then all bunches behave similarly. We obtain a $2^d$-ary tree. In this tree, each node corresponds to a rectangular region of $[0,1]^d$, and its children are the result of its split into $2^d$ subregions. The points come uniformly at random, and hence the probability that a region is hit is its area. Figure 5.11 illustrates the way we turn the question about the ratio into the weighted height of some tree.

The area of a rectangle is the product of $[0,1]$-uniform random variables determining the splits, and the ratio $L_1/L_2$ is the ratio of two products of some of these random variables. More precisely, let $U_i$, $1 \le i \le d$ be i.i.d. $[0,1]$-uniform random variables. Taking the logarithms of the areas and of the ratios, we see that the increments are distributed like

$$E = -\sum_{i=1}^{d} \log U_i \qquad \text{and} \qquad Z = -\log U_1 + \log U_2.$$

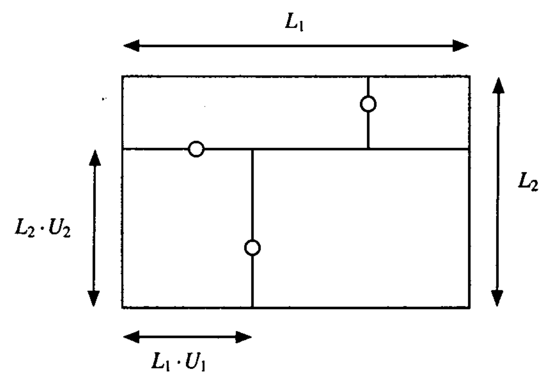


***Figure 5.11:*** *The way a node is interpreted when $d = 2$.*

To use Theorem 5.1, we need to compute $\Lambda^\star$ associated with $X = (Z, E)$. We start with the moment generating function: for any real numbers $\lambda$ and $\mu$, by independence,

$$\mathbf{E}\left[e^{\lambda Z + \mu E}\right] = \mathbf{E}\left[U_1^{-\lambda-\mu} \cdot U_2^{\lambda-\mu} \prod_{i=3}^{d} U_i^{-\mu}\right] = \mathbf{E}\left[U_1^{-\lambda-\mu}\right] \cdot \mathbf{E}\left[U_2^{\lambda-\mu}\right] \cdot \prod_{i=3}^{d} \mathbf{E}\left[U_i^{-\mu}\right].$$

As a consequence,

$$\mathbf{E}\left[e^{\lambda Z + \mu E}\right] = \begin{cases} \infty & \text{if } \lambda \geq 1-\mu \text{ or } \lambda \leq \mu - 1, \\ \dfrac{1}{(1-\mu)^2 - \lambda^2} \cdot \dfrac{1}{(1-\mu)^{d-2}} & \text{otherwise.} \end{cases}$$

It follows that

$$\Lambda(\lambda, \mu) = \begin{cases} \infty & \text{if } \lambda \geq 1-\mu \text{ or } \lambda \leq \mu - 1, \\ -\log\left((1-\mu)^2 - \lambda^2\right) - (d-2)\log(1-\mu) & \text{otherwise.} \end{cases}$$

So $\mathcal{D}_\Lambda = \{(\lambda, \mu) : \lambda < 1 - \mu, \lambda > \mu - 1\}$. To compute $\Lambda^\star$, we find the maximum of $(\lambda, \mu) \mapsto \lambda\alpha + \mu\rho - \Lambda(\lambda, \mu)$, which is achieved for $\lambda$ and $\mu$ such that

$$\begin{cases} \alpha = \dfrac{\partial \Lambda(\lambda, \mu)}{\partial \lambda} = \dfrac{2\lambda}{(1-\mu)^2 - \lambda^2} & \text{and} \\ \rho = \dfrac{\partial \Lambda(\lambda, \mu)}{\partial \mu} = \dfrac{2(1-\mu)}{(1-\mu)^2 - \lambda^2} + \dfrac{d-2}{1-\mu}, \end{cases} \tag{5.12}$$

if such a point exists. For $d \geq 2$, such a point does exist and the system above has solution

$$\begin{cases} \mu = 1 - \dfrac{\rho(d-1) \pm \sqrt{\rho^2 + \alpha^2 d(d-2)}}{\rho^2 - \alpha^2} \\ \lambda = \dfrac{\alpha}{\rho^2 - \alpha^2} \cdot \dfrac{2\rho(\mu-1)(d-1) + d(d-2)}{\rho(\mu-1) + d - 2}. \end{cases} \tag{5.13}$$

Observe that the mere fact that this is a solution of (5.12) ensures that $(\lambda, \mu) \in \mathcal{D}_\Lambda$. Hence, we have $\mathcal{D}_{\Lambda^\star} = \mathbb{R}^2$, and by Theorem 5.1, $c$ is the maximum of $\alpha/\rho$ with $(\alpha, \rho)$ in the set

$$\Psi(d \log 2) = \{(\alpha, \rho) : \lambda\alpha + \mu\rho + \log\left((1-\mu)^2 - \lambda^2\right) + (d-2)\log(1-\mu) \leq d \log 2\},$$

where $\lambda$ and $\mu$ are defined by (5.13). Since this only accounts for the levels whose depths are $0 \bmod d$, this gives only a lower bound on the actual weighted depth of the tree. However, one can find a matching upper bound easily. Indeed, to account for the levels $1 \bmod d$, it suffices to group the levels starting at level 1. Doing this, the distribution for $E$ and $Z$ remains unchanged, but the ratio $L_1/L_2$ is now off by one single multiplicative factor of $1/U$. It follows immediately that the weighted height on the levels $1 \bmod d$ is also $c \log n$, which finishes the proof of Theorem 5.8. □

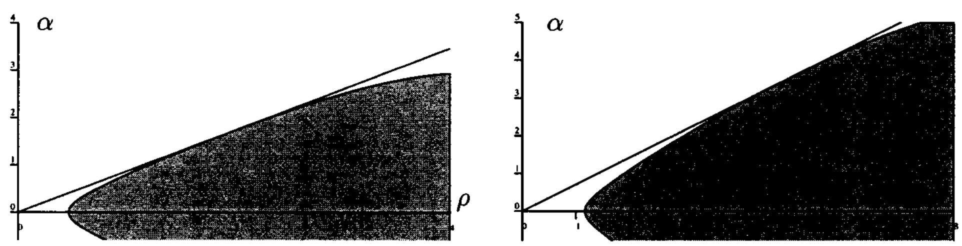


***Figure 5.12:*** *The sets $\{(\rho, \alpha) : \Lambda^\star(\alpha, \rho) \leq d \log 2\}$ together with the lines of maximum slopes for the maximum ratio of two dimensions of a cell in a k-d tree in $\mathbb{R}^2$ and $\mathbb{R}^5$.*

**Corollary 5.1.** *For $d = 2$, we have $c_2 = \sqrt{3}/2$.*

*Proof.* If $d = 2$, $\mu$ and $\lambda$ simplify and we have

$$\mu = 1 - \frac{2\rho}{\rho^2 - \alpha^2} \qquad \text{and} \qquad \lambda = \frac{2\alpha}{\rho^2 - \alpha^2}.$$

The condition $\mu - 1 < -|\lambda|$ is equivalent to $\rho > |\alpha|$, so the set to consider is $\{(\alpha, \rho) : \rho > |\alpha|\}$. It follows that

$$\Lambda^\star(\alpha, \rho) = \frac{2\alpha^2}{\rho^2 - \alpha^2} + \rho - \frac{2\rho^2}{\rho^2 - \alpha^2} + \log\left(\frac{2\lambda}{\alpha}\right) = \rho - 2 + 2\log 2 - \log(\rho^2 - \alpha^2).$$

Therefore, we need to find the maximum value of $\alpha/\rho$ subject to $\rho - 2 \le \log(\rho^2 - \alpha^2)$. The optimum is clearly obtained on the boundary of the set, i.e., for $\rho - 2 = \log(\rho^2 - \alpha^2)$. Then, we have

$$\frac{\alpha}{\rho} = \sqrt{1 - \frac{e^{\rho-2}}{\rho^2}},$$

which is maximum when the derivative vanishes:

$$\frac{d}{d\rho}\left(\frac{\alpha}{\rho}\right) = \frac{-e^{\rho-2}\left(\frac{1}{\rho^2} - \frac{2}{\rho^3}\right)}{2\sqrt{1 - \frac{e^{\rho-2}}{\rho^2}}} = 0.$$

This happens when $\rho = 2$ and then $\alpha/\rho = \sqrt{3}/2$. (Note that $\rho > |\alpha|$.) □

**Remark.** We have $\lim_{d\to\infty} c_d = 1/2$. Indeed, the optimal point is at $\rho = d$. Using $\alpha \sim cd$, we have

$$1 - \mu = \frac{1}{1-c} + o(1) \qquad \text{and} \qquad \lambda = \frac{1}{1-c} + o(1).$$

So,

$$\Lambda^\star(\alpha, \rho) = \frac{cd}{1-c} + d - \frac{d}{1-c} - d\log(1-c) + o(d).$$

Finally, if $c = 1/2$, we have $\Lambda^\star(\alpha, \rho) = d\log 2 + o(d)$.

### 5.7.13 Skinny cells in relaxed $k$-d trees

The model of $k$-d trees described above is a bit constrained due to the cyclical way in which the components of a vector are used as keys. In particular, $k$-d trees are data structures that are mostly static: they are built *once*, and then used to perform multiple queries on the *same* data. To cope with the issue of updating $k$-dimensional search structures, Duch and Martínez (2002) introduced a randomized data structure that is similar to $k$-d trees, but that does not suffer the same constraints. The symmetry is reintroduced by chosing the index of the component used as a key at random when a node is inserted in the structure. This tree structure is naturally called relaxed $k$-d tree. The structure leads to easy to present update algorithms, but it is not known whether the structure is indeed efficient.

**Theorem 5.9.** *Let $T_n$ be a relaxed k-d tree built from n i.i.d. $[0,1]^2$-uniform random points. Let*

$$R_n = \max\left\{\frac{L_1(u)}{L_2(u)} : u \in T_n\right\}.$$

*Then $R_n = n^{1+o(1)}$ in probability, as $n \to \infty$.*

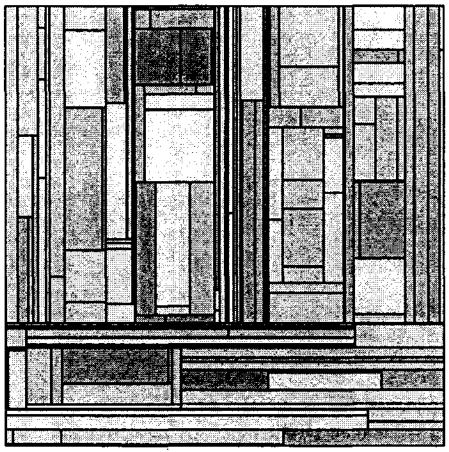

*Figure 5.13: A (randomly generated) relaxed k-d tree on 150 uniformly distributed points in $[0,1]^2$. One can notice in at the first glance that the cells look skinnier than those shown in Figure 5.10.*

**Remark.** The cells of 2-dimensional relaxed $k$-d trees are skinnier than those of $k$-d trees. This explains why partial match queries are more costly for relaxed $k$-d trees (Duch and Martínez, 2002; Duch, 2004) than for $k$-d trees (Flajolet and Puech, 1986).

*Proof of Theorem 5.9.* Consider a cell that does not contain any data point. In the tree, it corresponds to an external node $u$. A new incoming point falls in this cell with probability $L_1(u) \cdot L_2(u)$. If this happens, two new cells are created. Clearly, the cell gets divided uniformly. Let $U$ be a $[0,1]$-uniform random variable. Then, if the number $N_u$ of nodes contained in the subtree rooted at $u$ is $n$, the sizes of the subcells are distributed as a multinomial$(n-1, U, 1-U)$ random vector.

As in the case of $k$-d trees, the ratio $L_1/L_2$ is either multiplied or divided by $U$. Each of this cases happens with probability $1/2$ at every split, so with the additive formalism, the increase in $\log(L_1/L_2)$ is

$$Z(U) = \begin{cases} -\log U & \text{w.p.} \quad 1/2 \\ \log U & \text{w.p.} \quad 1/2. \end{cases}$$

Again, $\mathcal{X}^n \to ((Z(U), -\log U), (Z(1-U), -\log(1-U)))$ almost surely as $n \to \infty$. Hence we have $X = (Z(U), -\log U)$, and for $\lambda, \mu \in \mathbb{R}$,

$$\mathbf{E}\left[e^{\lambda Z + \mu E}\right] = \frac{1}{2}\mathbf{E}\left[U^{-\lambda-\mu}\right] + \frac{1}{2}\mathbf{E}\left[U^{\lambda-\mu}\right].$$

Therefore, we have

$$\mathbf{E}\left[e^{\lambda Z + \mu E}\right] = \begin{cases} \infty & \text{if } \lambda \geq 1-\mu \text{ or } \lambda \leq \mu - 1 \\ \dfrac{1-\mu}{(1-\mu)^2 - \lambda^2} & \text{otherwise.} \end{cases}$$

It follows that

$$\Lambda(\lambda, \mu) = \begin{cases} \infty & \text{if } \lambda \geq 1-\mu \text{ or } \lambda \leq \mu - 1 \\ \log(1-\mu) - \log\left((1-\mu)^2 - \lambda^2\right) & \text{otherwise.} \end{cases}$$

The maximum of $(\lambda, \mu) \mapsto \lambda\alpha + \mu\rho - \Lambda(\lambda, \mu)$ is achieved for $\lambda$ and $\mu$, with $\mu - 1 \leq \lambda \leq 1 - \mu$, satisfying

$$\alpha = \frac{\partial\Lambda(\lambda,\mu)}{\partial\lambda} = \frac{2\lambda}{(1-\mu)^2 - \lambda^2} \quad \text{and} \quad \rho = \frac{\partial\Lambda(\lambda,\mu)}{\partial\mu} = \frac{2(1-\mu)}{(1-\mu)^2-\lambda^2} - \frac{1}{1-\mu},$$

if such a point exists. This implies in particular that

$$\rho^2 - \alpha^2 = \frac{4(1-\mu)^2}{((1-\mu)^2-\lambda^2)^2} + \frac{1}{(1-\mu)^2} - \frac{4}{(1-\mu)^2-\lambda^2} - \frac{4\lambda^2}{((1-\mu)^2-\lambda^2)^2} = \frac{1}{(1-\mu)^2} \geq 0,$$

for $\mu - 1 \leq \lambda \leq 1 - \mu$. Then, provided $|\alpha| < |\rho|$, the solution is given by

$$\mu = 1 - \frac{1}{\sqrt{\rho^2 - \alpha^2}} \qquad \text{and} \qquad \lambda = \frac{\alpha}{\rho^2 - \alpha^2} \cdot \frac{1}{1 + (1-\mu)\rho}. \tag{5.14}$$

If $|\alpha| \geq |\rho|$, then $\Lambda^*(\alpha, \rho) = \infty$. Indeed, assume that $\alpha = \rho + \delta$, for some $\delta > 0$ (a symmetric argument holds when $\alpha = -\rho - \delta$). Let $\epsilon > 0$, and write $\lambda_0 = 1 - \mu - \epsilon_0 \leq 1 - \mu$. Then,

$$\begin{aligned} \Lambda^*(\alpha, \rho) &\geq \lambda_0\alpha + \mu_0\rho - \Lambda(\lambda_0, \mu_0) \\ &= \lambda_0\delta + \rho - \rho\epsilon + \log 2 + \log\epsilon + O(1/\lambda_0) \to \infty \end{aligned}$$

as $\lambda_0 \to \infty$. It follows by Theorem 5.1 that $c$ is the maximum value of $\alpha/\rho$ in the set

$$\Psi(\log 2) = \{\lambda\alpha + \mu\rho - \log(1-\mu) + \log\left((1-\mu)^2 - \lambda^2\right) \leq \log 2\},$$

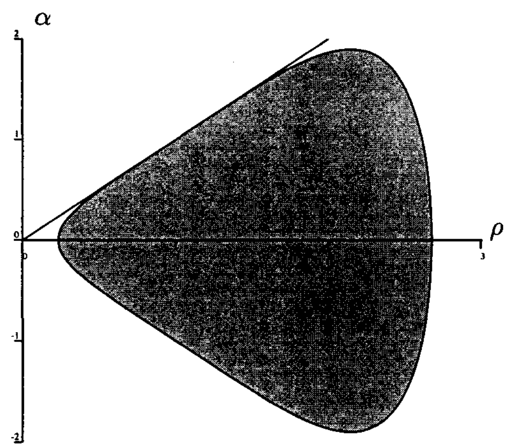


*Figure 5.14:* *The set* $\{(\rho, \alpha) : \Lambda^\star(\alpha, \rho) \leq \log 2\}$ *and the line of maximum slope for the maximum ratio of two dimensions of a cell in a relaxed k-d tree in* $\mathbb{R}^2$.

where $\lambda$ and $\mu$ are defined in (5.14).

Now, since $\mathcal{D}_{\Lambda^\star} \subset \{|\alpha| \leq |\rho|\}$, it is clear that $c \leq 1$, so we only need to prove that $c \geq 1$. In particular, it suffices to find points $(\alpha, \rho) \in \Psi(\log 2)$ with $\alpha/\rho$ arbitrarily close to 1. Because $\mu$ and hence $\Lambda^\star$ is not properly defined for $\alpha = \rho$, we consider $\Lambda^\star(1-\epsilon, 1)$ for $\epsilon \in (0,1)$. One can verify that, as $\epsilon \to 0$,

$$\Lambda^\star(1-\epsilon) = \log 2 - \sqrt{2\epsilon} + o\left(\sqrt{\epsilon}\right),$$

and therefore, $(1-\epsilon, 1) \in \Psi(\log 2)$ for $\epsilon$ small enough. This proves that $c \geq 1 - \epsilon$ for any small enough $\epsilon > 0$ and hence, by Theorem 5.1, that $\log R_n \sim \log n$ in probability as $n \to \infty$. □

### 5.7.14 $d$-ary pyramids

Allowing $Z = -\infty$ can be useful when one needs to exclude some tree paths in the definition of the height. Let us look at pyramids (Bhattacharya and Gastwirth, 1983; Gastwirth and Bhattacharya, 1984). These trees are built incrementally as follows: a $d$-ary pyramid of size 1 is a single node; given a $d$-ary pyramid of size $n$, pick a node $u$ uniformly at random among those that have degree at most $d-1$. The next node becomes a child of $u$. The height of a 2-ary pyramid has been studied by Mahmoud (1994). Biggins and Grey (1997) obtained it for $d \geq 2$.

**Theorem 5.10** (Mahmoud 1994; Biggins and Grey 1997). *The height $H_n$ of a d-ary pyramid of size n is $H_n \sim \log n/(\gamma\rho_0)$ in probability, as $n \to \infty$, where $\gamma$ is given by*

$$\sum_{i=1}^{d} \frac{1}{(1+\gamma)^i} = 1, \tag{5.15}$$

*and $\rho_0$ is defined as the smallest root of*

$$\rho \cdot \mu = \log\left(\sum_{i=1}^{d} \frac{1}{(1-\mu)^i}\right) \quad \text{where} \quad \rho = \frac{\sum_{i=1}^{d} i(1-\mu)^{-i-1}}{\sum_{i=1}^{d} (1-\mu)^{-i}}, \tag{5.16}$$

*where $\mu < 1$. Numerical values are given in Table 5.4.*

| | $d$ | | | |
|---|---|---|---|---|
| | 2 | 3 | 5 | 10 |
| $\gamma$ | 0.6180339880... | 0.8392867552... | 0.9659482366... | 0.9990186327... |
| $\rho_0$ | 0.4056580492... | 0.3759749401... | 0.3684055189... | 0.3678801695... |
| $(\gamma\rho_0)^{-1}$ | 3.988664818... | 3.169061969... | 2.810088635... | 2.720946695... |

***Table 5.4:*** *Some numerical values for the height of d-ary pyramids of effective size n. It is not surprising to observe that $(\gamma\rho_0) \to 1/e$ as $d \to \infty$, since the height of the random recursive tree is asymptotic to $e \log n$.*

We derive Theorem 5.10 using a our framework. Random recursive trees (Smythe and Mahmoud, 1995) are $\infty$-ary pyramids. A random recursive tree of size one consists of single node. A random recursive tree of size $n+1$ is built from one of size $n$ by picking a uniform random node $u$, and adding a new node as a child of $u$. Clearly, conditioning on the new node being a child of an unsaturated node $u$, $u$ is still uniform among the unsaturated nodes. Hence, one can see a $d$-ary pyramid as the subtree of a random recursive tree consisting only of the first $d$ children of any node (Figure 5.15).

This gives a simple way to obtain the height of $d$-ary pyramids: build a random recursive tree in which the first $d$ children of any node have an edge of weight 1 to their parent, and the others a weight of $-\infty$: $(Z_1, \dots, Z_d, Z_{d+1}, \dots) = (1, \dots, 1, -\infty, \dots)$. One can verify (see, e.g., Broutin and Devroye, 2006) that the (infinite) split vector

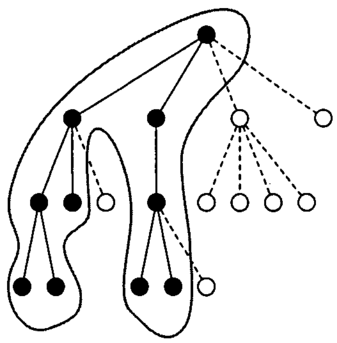

***Figure 5.15:*** *A 2-ary pyramid seen as the subtree of a random recursive tree consisting of the first two children of any node. The black vertices are part of the 2-ary pyramid.*

$(V_1, V_2, \ldots, V_i, \ldots)$ for a random recursive tree is distributed like

$$\left(U_1, (1-U_1)U_2, \ldots, \prod_{j=1}^{i-1}(1-U_j)U_i, \ldots\right),$$

where $\{U_i, i \geq 1\}$ is a family of i.i.d. $[0,1]$-uniform random variables. Since our result only holds for trees of bounded degree, we can rewrite the split vector by collecting the children with index greater than $d+1$ in a single "bin":

$$(V_1, V_2, \ldots, V_{d+1}) = \left(U_1, (1-U_1)U_2, \ldots, \prod_{j=1}^{d-1}(1-U_j)U_d, \prod_{j=1}^{d}(1-U_j)\right), \tag{5.17}$$

and $(Z_1, Z_2, \ldots, Z_{d+1}) = (1, \ldots, 1, -\infty)$. Write $E_k = -\log V_k$. The height is not affected by a random permutation of the children, so the random variable of interest is $X = (Z, E) = (Z_K, E_K)$, where $K$ be taken uniformly at random in $\{1, \ldots, d+1\}$. Then, according to the definition of $\Lambda$, we have that for all $\lambda$ and $\mu$ real numbers,

$$\Lambda(\lambda, \mu) = \log \mathbf{E}\left[\, e^{\lambda + \mu E_K} \;\middle|\; K \leq d\right] + \log d - \log(d+1).$$

Using the definition (5.17) for the split vector $(V_1, \ldots, V_{d+1})$, we find that,

$$\Lambda(\lambda, \mu) = \lambda + \log\left(\sum_{i=1}^{d} \mathbf{E}\left[U_1^{-\mu}\right]^i\right) - \log(d+1),$$

and therefore,

$$\Lambda(\lambda, \mu) = \begin{cases} \infty & \text{if } \mu \geq 1 \\ \lambda + \log\left(\sum_{i=1}^{d}(1-\mu)^{-i}\right) - \log(d+1) & \text{otherwise.} \end{cases}$$

We have $\mathcal{D}_\Lambda = \{(\lambda, \mu) : \mu < 1\}$. It follows that the optimum value for $\mu = \mu(\rho)$ is obtained for

$$\alpha = \frac{\partial\Lambda(\lambda,\mu)}{\partial\lambda} = 1 \qquad \text{and} \qquad \rho = \frac{\partial\Lambda(\lambda,\mu)}{\partial\mu} = \frac{\sum_{i=1}^{d} i(1-\mu)^{-i-1}}{\sum_{i=1}^{d}(1-\mu)^{-i}}. \tag{5.18}$$

The function $f$ defined on $(0,\infty)$ by

$$x \mapsto \sum_{i=1}^{d} i x^{-i-1} \Big/ \sum_{i=1}^{d} x^{-i}$$

is continuous, $\lim_{x\to\infty} f(x) = 0$ and $\lim_{x\downarrow 0} f(x) = \infty$. Therefore, (5.18) admits a well-defined solution $\mu(\rho)$ for all $\rho > 0$. By Theorem 5.2, one needs a rescaling factor $\gamma$ to express the height of $d$-ary pyramids of size $n$. The constant $\gamma$ is given by

$$\sum_{i=1}^{d} \frac{1}{(1+\gamma)^i} = 1.$$

Then, the height of a $d$-ary pyramid is asymptotic to $\log n/(\gamma\rho_0)$, where $\rho_0$ satisfies

$$\lambda + \mu(\rho_0)\rho_0 - \Lambda(\lambda, \mu(\rho_0)) = \log(d+1),$$

which proves (5.16) since $\lambda$ cancels.

*Chapter 6*

# Weighted height of tries

*We define a general model for weighted random tries. We analyze the weighted heights of such trees, using a link between the profile of a "core" of the trie. In particular, we apply our main result to the worst-case search time in trees introduced by de la Briandais (1959) and the ternary search trees of Bentley and Sedgewick (1997). The chapter is based on Broutin and Devroye (2007c) and Broutin and Devroye (2007b).*

*Everything you see, I owe to spaghetti.*
– Sophia Loren

## Contents

## 6.1 Introduction

In this chapter, we are interested in tries. We have introduced tries in section 1.4.5. For now, recall simply that they are tree-like data-structures used to store strings. Here, the alphabet is $\{1,\dots,d\}$. The strings are stored in the leaves, which may each contain one or more of them. We assume that the sequences are built using a memoryless source: each string is an infinite sequence of i.i.d. symbols distributed like $A \in \{1,\dots,d\}$, where $\mathbf{P}\{A=i\}=p_i$, $1 \le i \le d$. Also, without loss of generality $p_1 \ge p_2 \ge \cdots \ge p_d > 0$. A useful quantity under this model is the probability that $b$ independent characters are identical

$$Q(b)=\sum_{i=1}^{d} p_i^b. \tag{6.1}$$

It is well known that the height $H_n$ of a trie built from $n$ independent such sequences satisfies (Régnier, 1981; Devroye, 1984; Pittel, 1985; Szpankowski, 1991, 2001)

$$\frac{H_n}{\log n} \xrightarrow[n\to\infty]{} \frac{2}{-\log Q(2)} \quad \text{in probability.} \tag{6.2}$$

This holds for ordinary tries, i.e., if every leaf contains only one string. If the leaves can store up to $b$ sequences, the tree is called a $b$-trie and its height $H_{n,b}$ is such that

$$\frac{H_n}{\log n} \xrightarrow[n\to\infty]{} \frac{b+1}{-\log Q(b+1)} \quad \text{in probability.}$$

Park, Hwang, Nicodème, and Szpankowski (2006) have recently reproved (6.2) via the profile of the tree (number of nodes at each level). The results of this chapter are proved using a similar approach based on the profile.

The trie is only an *abstract data structure*, that is, it does not specify the implementation (see Clément et al., 1998, 2001). Depending on the implementation, the worst-case search time and the height of the trie may be different. This happens in particular when the implementation of a node relies on a linked-list or a search tree (de la Briandais, 1959; Bentley and Sedgewick, 1997) instead of an array. Our aim in this chapter is (1) to make a link between the worst-case search time and the *weighted height* of a tree that would hold for many implementations, and (2) to characterize the height, and hence the worst-case search time of these data structures.

The chapter is organized as follows. In Section 6.2, we describe a model of random weighted trie and state our main result concerning the weighted height of such trees. The proof of the main theorem is based on an analysis of the internal structure of the trie, and the notion of a *core* (see Broutin and Devroye, 2007a). We describe the core of the weighted trie in section 6.3, and the behavior of the trees hanging from the core in Section 6.4. The properties are then used in section 6.5 to prove Theorem 6.1. Finally, we give some applications in Section 6.6. In particular, we show that the heights of the trees of de la Briandais (1959), and the ternary search trees of Bentley and Sedgewick (1997) follow from Theorem 6.1.

## 6.2 A model of random tries

Consider the distribution $\{p_1, \dots, p_d\}$ over a finite alphabet $\mathcal{A} = \{1, \dots, d\}$. We assume without loss of generality that $1 > p_1 \geq p_2 \geq \cdots \geq p_d > 0$. We are given $n$ independent infinite sequences of i.i.d. characters of $\mathcal{A}$ generated using $\{p_1, \dots, p_d\}$. Let $T_\infty$ be an infinite $d$-ary position tree. Let $b \geq 1$ be a natural number.

THE SHAPE OF THE TRIE. Each string defines an infinite path in $T_\infty$. Let the cardinality $N_u$ of a node $u \in T_\infty$ be the number of strings whose path in $T_\infty$ intersect $u$. Then, for a natural number $b \geq 1$, the $b$-trie $T_{n,b}$ is constructed by pruning all the edges down any node of cardinality at most $b$. The sequences are distinct with probability one, and the strings define distinct paths in $T_\infty$. Therefore, the trie $T_{n,b}$

is almost surely finite. The tree $T_{n,b}$ constitutes the *shape* of the weighted trie. We define $E_i = -\log p_i$. For the edge $e$ between $u$ and its $i$-th child we let $p_e = p_i$ and $E_e = -\log p_e$.

There are $2^d$ *types* of nodes, each type being characteristic of the *branching structure* of the node. The branching structure of every node $u \in T_n$ is described by a $d$-vector $\tau_u$: if $u_1, \ldots, u_d$ are the $d$ children of $u$, then we define

$$\tau_u = (\mathbf{1}[N_{u_i} \geq 1], \mathbf{1}[N_{u_2} \geq 1], \ldots, \mathbf{1}[N_{u_d} \geq 1]) .$$

The vector $\tau_u$ indicates which one of the $d$ edges down $u$ are part of some path in $T_{n,1}$.

As in Chapter 5, we consider random tries that may be built using an embedding. Our construction emphasizes an underlying structure consisting of independent random variables. However, in the coupled tries built from the embedding, the random variables are dependent in general because of the construction process. Observe that our embedding is only one way to build tries with the desired distribution. We will show in Section 6.6 that many tries of interest are covered by this model.

THE WEIGHTS. We now describe the way in which the weights are assigned. Consider a sequence of random vectors $\{\mathcal{Z}^\tau, \tau \in \{0,1\}^d\}$, where $\mathcal{Z}^\tau = (Z_1^\tau, \ldots, Z_d^\tau)$. For a fixed type $\tau \in \{0,1\}^d$, the components $Z_1^\tau, \ldots, Z_d^\tau$ of $\mathcal{Z}^\tau$ may be dependent. We assume that for all $\tau \in \{0,1\}^d$, $\mathcal{Z}^\tau$ has non-negative components and is bounded. Each node of $T_\infty$ is assigned an independent copy of the whole sequence. The weights are then associated with the edges of $T_\infty$ based on the types of nodes they link. Consider a node $u \in T_\infty$, and its sequence $\{\mathcal{Z}^\tau, \tau \in \{0,1\}^d\}$. The edge $e_i$ between $u$ and its $i$-th child in $T_\infty$ is given the weight

$$Z_{e_i} = Z_i^{\tau_u} = \sum_{\tau \in \{0,1\}^d} Z_i^\tau \cdot \mathbf{1}[\tau_u = \tau].$$

We use the notations $Z_i^\tau$ and $Z_e$ interchangeably. It should always be clear whether a subscript refers to an index or an edge. Let $\pi(u)$ be the set of edges on the path from $u$ up to the root in $T_\infty$. The *weighted depth* of a node $u$ is defined by $D_u = \sum_{e \in \pi(u)} Z_e$.

Observe that weighted depths are associated to every node in $T_\infty$. We are interested in the weighted height of $T_{n,b}$ defined by

$$H_{n,b} = \max\{D_u : u \in T_{n,b}\}.$$

Surprisingly, the first asymptotic term of $H_{n,b}$ depends on four parameters only:

- the capacity $b$ of the leaves,
- the distribution $\{p_1, \ldots, p_d\}$,
- $\mathcal{Z}^c \stackrel{\text{def}}{=} \mathcal{Z}^{(1,\ldots,1)}$, and
- $\mathcal{Z}^s$ defined in the following way: there are $d$ permutations of $(1, 0, \ldots, 0)$. Let $\sigma_i$ be the one with the one in position $i$. Then, we let $\mathcal{Z}^s = (Z_1^s, \ldots, Z_d^s)$ with $Z_i^s = Z_i^{\sigma_i}$.

In particular, the first order asymptotics of $H_{n,b}$ stay the same if we modify $\{\mathcal{Z}^\tau, \tau \in \{0,1\}^d\}$ in such a way that both $\mathcal{Z}^c$ and $\mathcal{Z}^s$ remain unchanged. This is easier understood by thinking of the structure of the shape of a trie.

THE CORE OF A TRIE. The profile of a trie can be explained by distinguishing a so-called *core*, that constitutes the bulk of the trie, and *spaghetti*-like trees hanging from the core (Broutin and Devroye, 2007a). The core of the trie is defined to be the set of nodes $u \in T_\infty$ for which $N_u \geq m(n)$, for $m(n) \to \infty$ and $m(n) = o(\log n)$. The core is denoted by $\mathcal{C}$. The spaghettis are the trees remaining when pulling out all the nodes of the core. It is shown by Broutin and Devroye (2007a) that the core is very stable and is barely affected by the choice of the sequence $m(n)$. Also, since $m \to \infty$, we expect a node in the core to be of type $\tau = (1, \ldots, 1)$ with probability $1 - o(1)$. As a consequence, in a weighted trie, the distribution of weights in the core should be closely approximated by $\mathcal{Z}^c$ (where the superscript stands for "core").

HANGING SPAGHETTIS. Let $\partial\mathcal{C}$ be the node boundary of the core $\mathcal{C}$, that is, the set of nodes in $T_\infty \setminus \mathcal{C}$ with a parent in $\mathcal{C}$. The trees rooted at every node in $\partial\mathcal{C}$ contribute a large amount to the height (for instance, half of it in a symmetric trie). We call these trees the *spaghettis.* The spaghettis lie in the part of the trie where the nodes do not have $d$ children any more: the types of the nodes may take all the values in $\{0,1\}^d$. However, the weighted height of a *long* spaghetti is close to the weighted height it would have if we discard the nodes that are *truly* branching. To see this, observe that the nodes in $\partial\mathcal{C}$ have cardinality at most $m(n) = o(\log n)$, and each spaghetti stores at most $m(n)$ sequences. Each time the type $\tau$ is not a permutation of $(1,0,\dots,0)$, at least one string is put aside from the longest path. This can happen at most $o(\log n)$ times, and hence the heights with and without the branching nodes differ by at most $o(\log n)$. If the weighted height is $\Theta(\log n)$, as is the case for the highest ones, the difference is negligible. This explains why $\mathcal{Z}^s$ (the superscript stands for "spaghetti") only matters in the weighted heights of spaghettis.

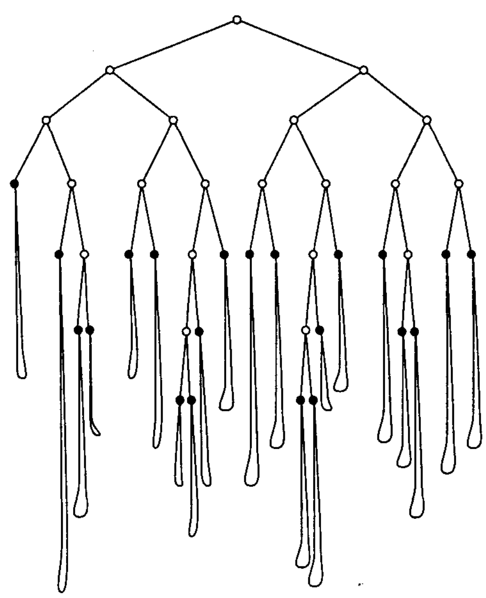

***Figure 6.1:*** *The structure of a trie: the bulk is the core. Some spaghetti-like trees hang down from the core. Both the core and the spaghettis contribute significantly to the height of the trie. Observe also that the height may not be explained by a spaghetti born at one of the deepest nodes of the core. This latter fact will become clear later.*

Both the core and the spaghettis contribute significantly to the height of a weighted trie. By figuring out what the core looks like, we can determine *when* the spaghettis take over. Roughly speaking, we then know if an edge's weight can be approximated by a component of $\mathcal{Z}^c$ or rather $\mathcal{Z}^s$. The shape of the core is the very question addressed by Broutin and Devroye (2007a) in the unweighted case, i.e., with all the weights equal to one. We shall rely on similar ideas here. The arguments are based on an analysis of the *logarithmic profile*

$$\phi(\alpha, t) = \lim_{n\to\infty} \frac{\log \mathbf{E} P_m(t \log n, \alpha \log n)}{\log n} \qquad \forall t, \alpha > 0, \tag{6.3}$$

where $P_m(k, h)$ denotes the number of nodes $u$, $k$ levels away from the root with $N_u \geq m(n)$ and $D_u \geq h$. Let $Z^b$ have the following distribution:

$$Z^b = \begin{cases} Z_A^s & \text{w.p.} \quad Q(b+1) \\ -\infty & \text{otherwise,} \end{cases} \tag{6.4}$$

where $A \in \{1, \ldots, d\}$ is a character generated by the memoryless source with probability distribution $\{p_1, \ldots, p_d\}$. The main result of this chapter is the following theorem.

**Theorem 6.1.** *Consider a weighted b-trie built from n independent sequences defined as above. Let $H_{n,b}$ be its weighted height. Let $\phi(\alpha, t)$ be the logarithmic weighted profile of the core of $T_{n,b}$. Let $\Lambda_b^\star$ be the rate function associated with $Z^b$. Let*

$$\gamma_b = \sup\left\{\gamma : \rho \cdot \Lambda_b^\star(\gamma/\rho) \leq 1, \gamma > 0, \rho > 0\right\}, \tag{6.5}$$

*and $c_b = \sup\{\alpha + \gamma_b \cdot \phi(\alpha, t) : \alpha, t > 0\}$. Then $H_{n,b} = c_b \log n + o(\log n)$ in probability, as $n \to \infty$.*

**Remarks.** (a) The contributions of the core and spaghettis are $\alpha$ and $\gamma_b \cdot \phi(\alpha, t)$, respectively. Both contributions are significant. Moreover, the joint of core and the spaghettis on the longest path, i.e, the level at which the longest path leaves the core is far from the bottom of the bottom of the core. See Figure 6.8.
(c) The definition of $c_b$ given makes it clear that $c_b > 0$ is well and uniquely defined. We will see later that $c_b < \infty$.

Before going further, we formalize our claim about the types of nodes that may influence the first order term of the height $H_{n,b}$. Here, the weights are irrelevant. Lemma 6.1 below is at the heart of the distinction between the core and the spaghettis. We prove:

**Lemma 6.1.** *Let $T_{n,b}$ be a random b-trie. There exists $\omega \to \infty$, as $n \to \infty$, such that on every path down the root,*

*(a) the number of nodes of the core not having $d$ children is $o(\log n)$ with probability $1 - n^{-\omega}$, and*

*(b) the number of nodes outside the core having at least two children is at most $m = o(\log n)$.*

*Proof.* The number of nodes with degree at least 2 in any spaghetti is at most $m(n) = o(\log n)$ and (b) follows. Therefore, we need only consider the portion of the paths that lie in the core and prove (a). We distinguish two regions of the core: the set of nodes $u$ such that $N_u \geq \log^2 n$, and the rest. The top of the core, consisting of nodes $u$ with $N_u \geq \log^2 n$, is very likely to be free of any node with less than $d$ children: in this region, with probability $1 - o(1)$, *all* the nodes have $d$ children. For any node $u$, we have

$$\mathbf{P}\left\{\tau_u \neq (1, \ldots, 1) \mid N_u \geq \log^2 n\right\} \leq d(1 - p_d)^{\log^2 n}.$$

Moreover, the number of such nodes is polynomial in $n$. Let $\mathscr{L}_k$ be the set of nodes at level $k$ in $T_\infty$. Indeed, at distance $k = \lceil \log_{1/p_1} n \rceil$ from the root, for $n$ large enough,

$$\begin{aligned}\mathbf{P}\left\{\exists u \in \mathscr{L}_k : N_u \geq \log^2 n\right\} &\leq d^{\log_{1/p_1} n + 1} \cdot \mathbf{P}\left\{\mathrm{Bin}(n, p_1^k) \geq \log^2 n\right\} \\ &\leq dn^{\log_{1/p_1} d} \cdot e^{-\frac{1}{2}\log^2 n},\end{aligned}$$

by Chernoff's bound. Therefore, by the union bound,

$$\begin{aligned}\mathbf{P}\left\{\exists u : N_u \geq \log^2 n, \tau_u \neq (1, \ldots, 1)\right\} &\leq dn^{\log_{1/p_1} d} \cdot \left(2(1 - p_1)^{\log^2 n} + e^{-\frac{1}{2}\log^2 n}\right) \\ &\leq n^{-\omega_1}, \end{aligned} \tag{6.6}$$

for some $\omega_1 \to \infty$ as $n \to \infty$.

There is also a number of layers of nodes $u$ with $m(n) \leq N_u < \log^2 n$. There are only $o(\log n)$ such layers. To see this, let $\nu = \nu(n) \to \infty$ to be chosen later, and look at a node $v$ lying $\lceil \frac{1}{\nu} \log n \rceil$ levels away from $u$ with $N_u \leq \log^2 n$. Then,

$$\mathbf{P}\left\{N_v \geq m(n)\right\} \leq \mathbf{P}\left\{\mathrm{Bin}(\log^2 n, p_1^{\frac{1}{\nu}\log n}) \geq m\right\}. \tag{6.7}$$

The expected value of the binomial random variable above is

$$\ell = \log^2 n \cdot p_1^{\frac{1}{\nu}\log n} = \log^2 n \cdot n^{\frac{1}{\nu}\log p_1} \xrightarrow[n\to\infty]{} 0, \tag{6.8}$$

for $\nu = o(\log n/\log\log n)$. In particular, for $n$ large enough, $\ell \leq m/2$. By the Chernoff bound for binomial random variables (see, e.g., Janson et al., 2000),

$$\mathbf{P}\left\{\mathrm{Bin}(\log^2 n, p_1^{\frac{1}{\nu}\log n}) \geq m\right\} \leq \exp\left(-\ell\varphi\left(\frac{m}{2\ell}\right)\right), \tag{6.9}$$

where $\varphi(x) = (1+x)\log(1+x) - x$. Using (6.8), we see that, as $n \to \infty$,

$$\begin{aligned}
\ell\varphi\left(\frac{m}{2\ell}\right) &= \left(\ell + \frac{m}{2}\right)\log\left(1 + \frac{m}{2\ell}\right) - \frac{m}{2} \\
&\sim \frac{m}{2}\cdot\log\left(\frac{m}{2\ell}\right) \\
&\sim \frac{m}{2}\log\left(\frac{m}{2}\right) - m\log\log n - \frac{m}{2\nu}\log p_1 \log n \\
&\sim \frac{m}{2\nu}\log\left(\frac{1}{p_1}\right)\log n,
\end{aligned}$$

for $\nu = o(\log n/\log\log n)$. We now choose $\nu$ such that, in addition, $\nu = o(m)$ so that, by (6.7) and (6.9), $\mathbf{P}\{N_v \geq m\}$ decreases faster than any polynomial in $n$. The number of potential nodes $v$ is polynomial in $n$ since they lie $O(\log n)$ away from the root. It follows that the maximum number of levels between a node $u$ with $N_u \leq \log^2 n$ and $v$ such that $N_v \leq m$ is $O(\frac{\log n}{\nu}) = o(\log n)$ with probability at least $1 - n^{-\omega_2}$, for some $\omega_2 \to \infty$ as $n \to \infty$. With (6.6), this proves the claim with $\omega = \min\{\omega_1, \omega_2\}/2$. □

## 6.3 The core of a weighted trie

### 6.3.1 Asymptotic behavior

Consider a weighted $b$-trie defined as in Section 6.2. We consider $m = m(n) \to \infty$ with $m(n) = o(\log n)$. Let $\mathscr{L}_k$ be the set of nodes $k$ levels away from the root in $T_\infty$. Let $P_m(k,h)$ be the number of nodes $u \in \mathscr{L}_k$ with $D_u \geq h$ and $N_u \geq m$. Since $m \to \infty$, for $n$ large enough, we have $m \geq b$ and

$$P_m(k,h) = \sum_{u\in\mathscr{L}_k} \mathbf{1}[N_u \geq m, D_u \geq h].$$

The first step in characterizing the profile is to study its expected value, we then use some concentration arguments.

The asymptotic properties of the expected profile are directly tied to large deviation theory (see Chapter 2). The random vector of interest here is $(Z^c, E) \stackrel{\text{def}}{=} (Z^c_K, -\log p_K)$, where $K$ is uniform in $\{1, \ldots, d\}$ and $\mathcal{Z}^c = (Z^c_1, \ldots, Z^c_d)$. Here, $\mathbf{P}\{Z^c = -\infty, E = +\infty\} = 0$ and for $\lambda, \mu \in \mathbb{R}$, the associated *generating function of the cumulants* is

$$\Lambda(\lambda, \mu) = \log \mathbf{E}\left[e^{\lambda Z^c + \mu E}\right].$$

Recall that the definition of the convex dual $\Lambda^\star$ of $\Lambda$: for $x, y \in \mathbb{R}$,

$$\Lambda^\star(x, y) = \sup_{\lambda, \mu}\{\lambda x + \mu y - \Lambda(\lambda, \mu)\},$$

(see Chapter 2), and

$$I(x, y) = \inf\{\Lambda^\star(x', y') : x' > x, y' < y\}.$$

**Theorem 6.2.** *Let $m = m(n) \to \infty$ with $m = o(\log n)$. Let $k \sim t \log n$ and $h \sim \alpha \log n$ for some positive constants $t$ and $\alpha$. Let*

$$\phi(\alpha, t) = t \log d - t \cdot I\left(\frac{\alpha}{t}, \frac{1}{t}\right). \tag{6.10}$$

*If $\phi(\alpha, t) > -\infty$, then $\mathbf{E}P_m(k, h) = n^{\phi(\alpha,t)+o(1)}$, as $n \to \infty$. Moreover, for any $\epsilon > 0$, there exists $n$ large enough that, uniformly in any compact subset of $\{(\alpha, t) : t > 0, \phi(\alpha, t) > -a\}$, for any $a > 0$,*

$$\mathbf{E}P_m(k, h) \leq n^{\phi(\alpha,t)+\epsilon}.$$

**Remarks.** (a) Observe that Theorem 6.2 justifies the definition of $\phi(\cdot, \cdot)$ in (6.3).
(b) The constraint that $m(n)$ is $o(\log n)$ is only used in the lower bound. However, the main reason why we choose $m = o(\log n)$ is for the spagghettis to contain each a number of nodes of degree at least two of order $o(\log n)$. The lower bound actually holds for $m$ as large as $n^{o(1)}$. This has no effect on the upper bound.

(c) Theorem 6.2 is actually an easy extension of the results in Chapter 5. Indeed, for the model treated there, only $\{(\alpha,t):\phi(\alpha,t)=0\}$ matters, and we did not bother computing the entire profile.

Unlike the profile of *unweighted* tries (Devroye, 2002, 2005; Park et al., 2006), that of *weighted* tries does not seem concentrated. However, it is log-concentrated in the sense of the following theorem.

**Theorem 6.3.** *Let $m=m(n)\to\infty$ as $n\to\infty$ such that $m=o(\log n)$. Let $k=\lfloor t\log n-3\log\log n\rfloor$ and $h=\alpha\log n$ for some positive constants $t$ and $\alpha$ such that $\phi(\alpha,t)>0$. Then, for all $\epsilon>0$, as $n\to\infty$,*

$$\mathbf{P}\left\{P_m(k,h)\le n^{\phi(t,\alpha)-\epsilon}\right\}\xrightarrow[n\to\infty]{}0.$$

*If the convergence of $h/\log n\to\alpha$ and $k/\log n\to t$, as $n\to\infty$ is uniform in a compact subset $C$ of $\{(\alpha,t):\alpha,t>0,\phi(\alpha,t)>-a\}$ for some $a>0$, for all $n$ large enough,*

$$\sup_{(\alpha,t)\in C}\mathbf{P}\left\{P_m(k,h)\ge n^{\phi(t,\alpha)+\epsilon}\right\}\le n^{-\epsilon/2}.$$

Before we proceed with the proof, we shall derive some useful properties of the logarithmic profile $\phi(\cdot,\cdot)$ and describe its geometry.

**Lemma 6.2.** *Let $\phi(\cdot,\cdot)$ be the logarithmic profile as defined in (6.10). Then,*
*(a) $\phi(\cdot,\cdot)$ is concave,*
*(b) for all $a\in\mathbb{R}$, the level set $\Delta_\phi(a)\stackrel{def}{=}\{(\alpha,t):\alpha,t>0,\phi(\alpha,t)>a\}$ is bounded, and*
*(c) $\phi(\cdot,\cdot)$ is continuous on $\Delta_\phi(a)$, for all $a\in\mathbb{R}$.*

*Proof.* We refer the reader to Chapter 2 for properties of $\Lambda^\star(\cdot,\cdot)$ and $I(\cdot,\cdot)$.
(a) Recall in particular that $\Lambda^\star(\cdot,\cdot)$ and $I(\cdot,\cdot)$ are convex. The functions $(t,\alpha)\mapsto\alpha/t$, and $(t,\alpha)\mapsto 1/t$ are convex as well. It follows that $(\alpha,t)\mapsto I(\alpha/t,1/t)$ is convex and the result follows.
(b) We show that $\{(\alpha,t):\alpha,t>0,\Lambda^\star(\alpha/t,1/t)\le\log d\}$ is bounded. This is based on an upper bound for $\Lambda(\lambda,\mu)$: for $\lambda>0$ and $\mu<0$

$$\Lambda(\lambda,\mu)=\log\mathbf{E}\left[e^{\lambda Z+\mu E}\right]\le\log\mathbf{E}\left[e^{\lambda\|Z\|_\infty-\mu\log p_d}\right]=\lambda\|Z\|_\infty-\mu\log p_d.$$

It follows that for all $\alpha, t$ such that $\alpha/t > \|Z\|_\infty$ or $1/t < -\log p_d$,

$$\Lambda^\star\left(\frac{\alpha}{t}, \frac{1}{t}\right) \geq \lambda\left(\frac{\alpha}{t} - \|Z\|_\infty\right) + \mu\left(\frac{1}{t} + \log p_d\right) \to \infty$$

by choice of $\lambda \to \infty$ and $\mu \to -\infty$. Therefore, since $\Lambda^\star$ is continuous where it is finite (Lemma 2.2), for all $a \in \mathbb{R}$

$$\Delta_\phi(a) = \{(\alpha, t) : \alpha, t > 0, \phi(\alpha, t) > a\} \subset \left(0, \frac{\|Z\|_\infty}{-\log p_d}\right] \times \left(0, \frac{1}{-\log p_d}\right],$$

which is bounded. Observe in particular that $\{(\alpha, t) : \alpha, t > 0, \phi(\alpha, t) > -\infty\}$ is bounded as well.

(c) This is straightforward from (a) and (b). □

**Example: asymmetric randomized list-tries.** Consider asymmetric tries on $\{1, 2\}$ with $p_1 = p > 1/2$ and $p_2 = q = 1 - p$. The node structure is implemented using a linked-list. A fair coin is flipped independently at each node to decide whether 1 or 2 is first in the list. Therefore, the vector $\mathcal{Z} = (Z_1, Z_2)$ of search-costs is such that $Z_1$ and $Z_2$ take values 1 or 2 with equal probability. The variables $E$ and $Z$ are independent. They are both linear functions of Bernoulli random variables (see Dembo and Zeitouni, 1998, section 4.2 on transformations of large deviation functions). If we write $f(x) = x\log x + (1 - x)\log(1 - x) + \log 2$, then $\Lambda^\star(\alpha, \rho) = \Lambda_Z^\star(\alpha) + \Lambda_E^\star(\rho)$, where

$$\Lambda_Z^\star(\alpha) = f(\alpha - 1) \qquad \text{and} \qquad \Lambda_E^\star(\rho) = f\left(\frac{\rho + \log p}{\log p - \log q}\right).$$

Here, $\Lambda^\star$ is of the form $f(x) + f(y)$. The change of coordinates in $\phi(\cdot, \cdot)$ implies that the level sets of $\phi$ are triangles. The logarithmic profile shown in Figure 6.5 is taken from this example.

The definition of $\phi(\alpha, t)$ involves the function $I(\cdot, \cdot)$ that is slightly more complicated than $\Lambda^\star(\cdot, \cdot)$. In general, it is easier to study

$$\phi_o(\alpha, t) = t\log d - \Lambda^\star\left(\frac{\alpha}{t}, \frac{1}{t}\right). \qquad (6.11)$$

The value of $\phi(\alpha,t)$ only differs from $\phi_o(\alpha,t)$ when $\alpha \le t\mathbf{E}Z$ and $t\mathbf{E}E \le 1$. For such values of $\alpha$ and $t$, $\phi(\alpha,t) = t\log d$.

**Remark.** Although we will not prove it, it is interesting to note that $\phi_o(\alpha,t)$ is the logarithmic profile counting the number of nodes at level $k \sim t\log n$ with $N_u$ *roughly* $m$ and $D_u$ *roughly* $h \sim \alpha\log n$. Proving this would require the equivalent of the Bahadur and Rao (1960) theorem for large deviations tail probabilities. See Broutin and Devroye (2007a) for a proof of this claim in the unweighted case.

Before we look at the two-dimensional version, we warm up and analyze the unweighted case.

THE UNWEIGHTED CASE. Here we assume $Z_e = 1$ almost surely for every edge $e$. In this paragraph only, we write $\Lambda^\star(t)$, $I(t)$ and $\phi(t)$ since the $\alpha$ variable is irrelevant. We have seen in Chapter 2 that $\Lambda^\star(t)$ is convex and looks like the function shown on Figure 6.2.

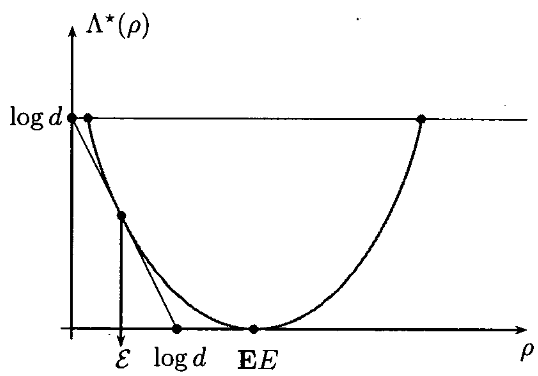


***Figure 6.2:*** *The function $\Lambda^\star$ corresponding to the distribution* $\{.9,.1\}$.

However, the logarithmic profile $\phi(t)$ is defined in terms of

$$I(\rho) = \inf\{\Lambda^\star(x) : x < \rho\} = \begin{cases} \Lambda^\star(\rho) & \text{if } \rho \le \mathbf{E}E \\ 0 & \text{otherwise.} \end{cases}$$

Finally, we have (see Figure 6.3),

$$\phi(t) = \begin{cases} t\log d - t\Lambda^\star(1/t) & \text{if } t \ge \frac{1}{\mathbf{E}E} \\ t\log d & \text{otherwise.} \end{cases}$$

So, $\phi(t)$ is linear to the left of $1/\mathbf{E}E$. Also, $\phi(t)$ is maximum at $t = 1/\mathcal{E}$, where $\mathcal{E} = -\sum_{i=1}^{d} p_i \cdot \log p_i$ is the entropy associated with the distribution. The following lemma gives the main properties of $\phi$.

**Lemma 6.3.** *We have the following:*
*(a)* $\Lambda^\star(\mathbf{E}E) = 0$ *and, for all* $t \in \mathcal{D}_{\Lambda^\star}$, $\Lambda^\star(t) \in [0, \ln d]$,
*(b) if* $(p_1, \ldots, p_d) \neq (1/d, \ldots, 1/d)$,

$$\left.\frac{\partial \phi(t)}{\partial t}\right|_{t=1/\mathbf{E}E} = \log d.$$

*Proof.* (a) This is clear since $E$ has just $d$ equiprobable masses, $-\log p_i$, $1 \leq i \leq d$. For all $\rho \in [-\log p_1, -\log p_d]$, $\Lambda^\star(\rho) \in [0, \log d]$. Outside this interval, $\Lambda^\star = +\infty$.
(b) If there exists an $i$ such that $p_i \neq 1/d$, $E$ is not degenerate and $\mathbf{E}E$ lies in the interior $\mathcal{D}_{\Lambda^\star}$. Then $\Lambda^\star$ is differentiable at $\mathbf{E}E$ and

$$\left.\frac{\partial \Lambda^\star(\rho)}{\partial \rho}\right|_{\rho=\mathbf{E}E} = 0.$$

It follows that

$$\left.\frac{\partial \phi(t)}{\partial t}\right|_{t=1/\mathbf{E}E} = \left.\frac{\partial \phi_o(t)}{\partial t}\right|_{t=1/\mathbf{E}E} = \log d + \frac{1}{t} \cdot \left.\frac{\partial \Lambda^\star(\rho)}{\partial \rho}\right|_{\rho=\mathbf{E}E} - \Lambda^\star(\mathbf{E}E) = \log d. \quad \square$$

The function $\phi(\cdot)$ can be characterized as the lower envelope of a set of lines. Let us define the generalized entropy function:

$$\mathcal{E}(b) = -\sum_{i=1}^{d} \log p_i \cdot p_i^b.$$

So, we have $\mathcal{E}(1) = \mathcal{E}$. The following lemma explains the geometry of $\phi(\cdot)$.

**Lemma 6.4.** *Let* $b > 0$. *Assume that* $(p_1, \ldots, p_d) \neq (1/d, \ldots, 1/d)$. *Let* $t_o = t_o(b) = Q(b)/\mathcal{E}(b)$. *The line of support of* $\phi$ *at* $t_o$ *has equation* $b + t \log Q(b)$. *Hence, we have*

$$\phi(t) = \inf_{b>0} \left\{ b - t \log \left( \frac{1}{Q(b)} \right) \right\}.$$

*In particular,* $\phi(1/\mathcal{E}) = 1$ *and the tangent of* $\phi(t)$ *at* $t = 1/\mathcal{E}$ *is horizontal.*

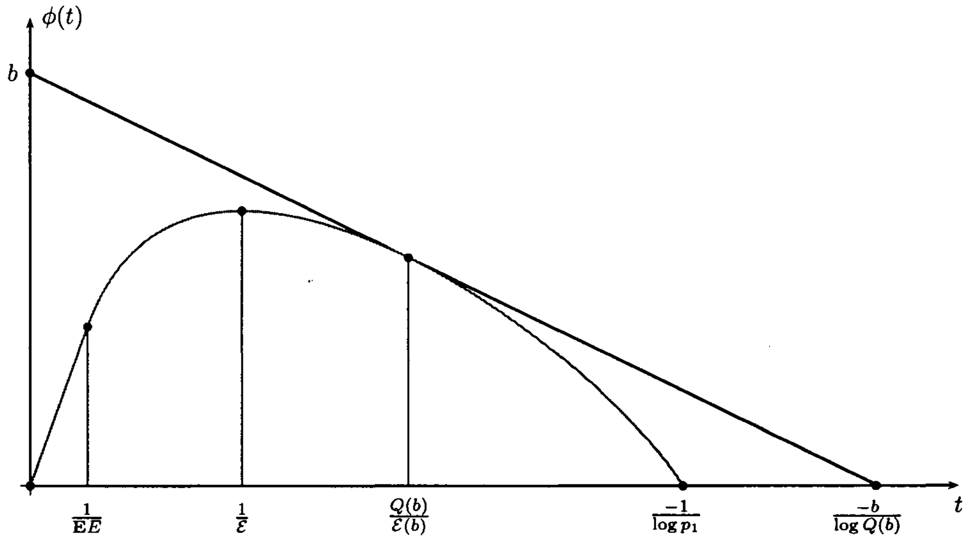


***Figure 6.3:*** *The logaritmic unweighted profile $\phi$ associated with the binary distribution $p_1 = .9$ and $p_2 = .1$.*

*Proof.* We only need to verify that the slope of $\phi(t)$ at $t_o$ is $\log Q(b)$. Consider first $t \geq 1/\mathbf{E}E$. Then $\Lambda^\star(1/t) = I(1/t)$, and we have $\phi(t) = t \log d - t\Lambda^\star(1/t)$. So,

$$\frac{\partial \phi(t)}{\partial t} = \log d - \Lambda^\star(1/t) + \frac{1}{t} \cdot \left.\frac{\partial \Lambda^\star(\rho)}{\partial \rho}\right|_{\rho=1/t}.$$

Also, $\Lambda^\star(\rho) = \lambda\rho - \Lambda(\lambda)$, where $\lambda = \lambda(\rho)$ is defined by

$$\rho = \frac{\partial \Lambda(\lambda)}{\partial \lambda}.$$

For $t = t_o$, we have $\lambda = \lambda(1/t_o) = -b$ and

$$\begin{aligned}
\left.\frac{\partial \phi(t)}{\partial t}\right|_{t=t_o} &= \log d - \Lambda^\star(1/t) + \frac{1}{t} \cdot \left.\frac{\partial \Lambda^\star(\rho)}{\partial \rho}\right|_{\rho=1/t_o} \\
&= \log d - \left(-b\frac{\mathcal{E}(b)}{Q(b)} - \log_d Q(b)\right) \\
&\quad + \frac{\mathcal{E}(b)}{Q(b)}\left(-b + \frac{\mathcal{E}(b)}{Q(b)}\left.\frac{\partial \lambda(\rho)}{\partial \rho}\right|_{\rho=1/t_o} - \frac{\mathcal{E}(b)}{Q(b)}\left.\frac{\partial \lambda(\rho)}{\partial \rho}\right|_{\rho=1/t_o}\right) \\
&= \log Q(b).
\end{aligned}$$

This is also Lemma 2.2.5 (c) of Dembo and Zeitouni (1998). Now, observe that the line of support of $\phi(\cdot)$ at $t_o$ hits the vertical axis at

$$\phi(t_o) - t_o \left.\frac{\partial \phi(t)}{\partial t}\right|_{t=t_o} = -\left.\frac{\partial \Lambda^\star(\rho)}{\partial \rho}\right|_{\rho=1/t_o} = b.$$

If, on the other hand $t \leq 1/\mathbf{E}E$, we have $\phi(t) = t \log d$. Since $\phi(\cdot)$ is concave, it is the lower envelope of its lines of support. This completes the proof. □

THE WEIGHTED CASE. In the general case, since $\phi(\alpha, t)$ is in terms of $I(\alpha/t, 1/t)$, one can picture $\phi(\cdot,\cdot)$ by considering diagonal slices $\phi(xt, t)$, when $x > 0$. In particular, $\phi(xt, t) = \phi(\mathbf{E}Zt, t)$ for all $x \leq \mathbf{E}Z$. Some of the main properties of $\Lambda^\star(\cdot,\cdot)$ and $\phi(\cdot,\cdot)$ are given in the next lemma. Typical curves are shown in Figures 6.4 and 6.5.

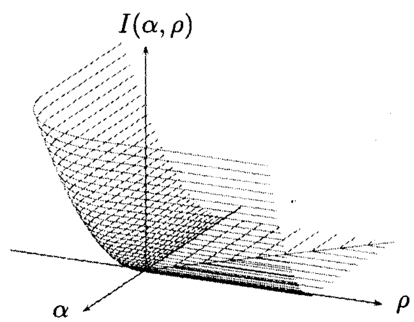


*Figure 6.4: A typical rate function $I(\cdot,\cdot)$ (the one from the example of gaussians in Chapter 2).*

**Lemma 6.5.** *For all $\alpha > 0$ and $t > 0$, $\phi(\alpha, t) \leq 1$.*

Although the result is clear if we take Theorem 6.2 for granted, we prove it from the first principles. Indeed, it is needed in the proof of our main result.

*Proof.* If $p_1 = \cdots = p_d = 1/d$, the result is clear. We now assume that $p_i \neq 1/d$, for some $i$, $1 \leq i \leq d$. In such a case, $\Lambda^\star$ is differentiable in $\mathcal{D}^o_{\Lambda^\star}$ which is not empty. Again, we consider the case $\phi(t) = \phi_o(t)$ first. Recall that, then,

$$\phi(\alpha, t) = \phi_o(\alpha, t) = t \log d - t\Lambda^\star\left(\frac{\alpha}{t}, \frac{1}{t}\right).$$

Also, for any $x, y$, $\Lambda^\star(x, y) = \lambda x + \mu y - \Lambda(\lambda, \mu)$ where $\lambda = \lambda(x, y)$ and $\mu = \mu(x, y)$ satisfy

$$x = \frac{\partial \Lambda(\lambda, \mu)}{\partial \lambda} \qquad \text{and} \qquad y = \frac{\partial \Lambda(\lambda, \mu)}{\partial \mu}. \tag{6.12}$$

For any $\alpha, t$, let $\lambda_o$ and $\mu_o$ be the values of $\lambda(x,y)$ and $\mu(x,y)$ at $(x,y) = (\frac{\alpha}{t}, \frac{1}{t})$. We have

$$\frac{\partial \phi_o(\alpha,t)}{\partial \alpha} = t \cdot \frac{1}{t} \cdot \left.\frac{\partial \Lambda^*(x,y)}{\partial x}\right|_{(x,y)=(\frac{\alpha}{t},\frac{1}{t})} = \lambda_o,$$

by the definition of $\Lambda^*$ and (6.12). Also, as for the partial derivative with respect to $t$:

$$\begin{aligned}
\frac{\partial \phi_o(\alpha,t)}{\partial t} &= \log d - \Lambda^*\left(\frac{\alpha}{t}, \frac{1}{t}\right) - t \cdot \frac{\partial \Lambda^*(\frac{\alpha}{t}, \frac{1}{t})}{\partial t} \\
&= \frac{\phi_o(\alpha,t)}{t} + \frac{\alpha}{t} \cdot \left.\frac{\partial \Lambda^*(x,y)}{\partial x}\right|_{(x,y)=(\frac{\alpha}{t},\frac{1}{t})} + \frac{1}{t} \cdot \left.\frac{\partial \Lambda^*(x,y)}{\partial y}\right|_{(x,y)=(\frac{\alpha}{t},\frac{1}{t})} \\
&= \frac{\phi_o(\alpha,t)}{t} + \frac{\alpha}{t} \cdot \lambda_o + \frac{1}{t} \cdot \mu_o.
\end{aligned}$$

Recall that

$$\Lambda(\lambda_o, \mu_o) = \frac{\alpha}{t}\lambda_o + \frac{\mu_o}{t} - \Lambda^*\left(\frac{\alpha}{t}, \frac{1}{t}\right).$$

Hence, we obtain

$$\frac{\partial \phi_o(\alpha,t)}{\partial t} = \frac{\log d}{t} + \frac{\Lambda(\lambda_o,\mu_o)}{t}.$$

At the maximum of $\phi_o(\alpha,t)$, the first partial derivatives with respect to $\alpha$ and $t$ both vanish. Hence, $\lambda_o = 0$ and $-\log d = \Lambda(0,\mu_o) = \log_d Q(-\mu_o)$, which implies $\mu_o = -1$. As consequence, for all $\alpha, t > 0$, such that $\phi(\alpha,t) = \phi_o(\alpha,t)$,

$$\phi(\alpha,t) = \phi_o(\alpha,t) = t\log d - t\Lambda^*\left(\frac{\alpha}{t}, \frac{1}{t}\right) \le -\mu_o = 1.$$

In the case when $\phi(\alpha,t) \neq \phi_o(\alpha,t)$, we have $t\mathbf{E}E \le 1$ and thus $t\log d \le \frac{\log d}{\mathbf{E}E}$. However, by Jensen's inequality ,

$$\mathbf{E}E = \frac{1}{d} \cdot \sum_{i=1}^{d} -\log p_i \ge -\log\left(\frac{1}{d}\sum_{i=1}^{d} p_i\right) = \log d.$$

Hence $t\log d \le 1$. This completes the proof. □

### 6.3.2 The expected logarithmic profile: Proof of Theorem 6.2

We first prove the upper bound of Theorem 6.2:

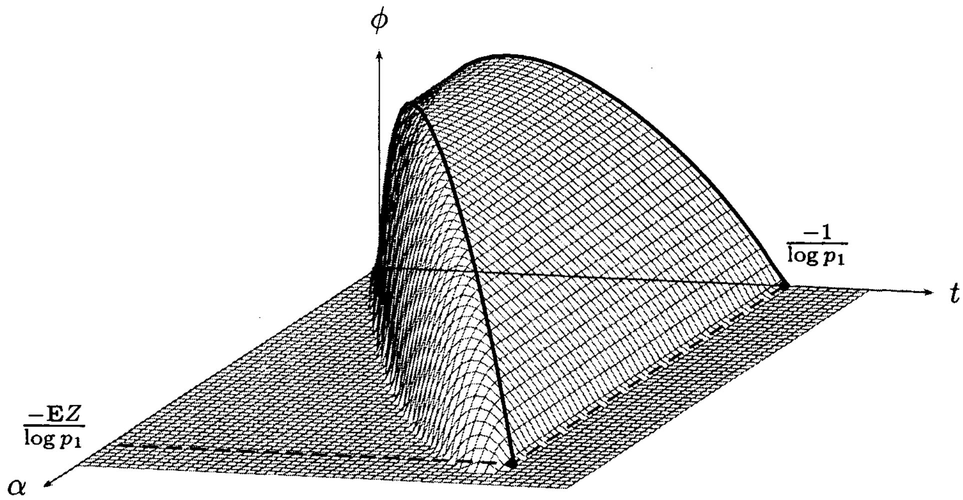


*Figure 6.5: A typical logarithmic profile. The thick black lines represent $\phi(0,t)$ and $\phi(t\mathbf{E}Z,t)$. For $t_o$ constant, $\phi(\alpha,t_o)$ is constant for $\alpha \in [0, t_o\mathbf{E}Z]$.*

**Lemma 6.6.** *Let $m = m(n) \to \infty$. Let $k \sim t\log n$ and $h \sim \alpha \log n$ for some positive constants $t$ and $\alpha$. Let $\phi(\alpha,t)$ be given by (6.10). If $\phi(\alpha,t) > -\infty$, then*

$$\mathbf{E}P_m(k,h) \le n^{\phi(\alpha,t)+o(1)},$$

*as $n \to \infty$. Moreover if the convergence of $(k/\log n, h/\log n)$ is uniform in a compact subset $\Gamma$ of $\Delta_\phi(-a) = \{(\alpha,t) : t > 0, \phi(\alpha,t) > -a\}$ for some $a > 0$, then for any $\epsilon > 0$, there exists $n$ large enough that*

$$\mathbf{E}P_m(k,h) \le n^{\phi(\alpha,t)+\epsilon}.$$

*uniformly in $\Gamma$.*

The ideas are similar to those we used in Chapter 4. The variables are not quite i.i.d. as in ideal trees. However, Lemma 6.1 gives us a good handle on the number of variables that are not of the expected type $(1,\ldots,1)$ in the core.

*Proof of Lemma 6.6.* It suffices to prove the result on $\Delta(-a)$ for an arbitrary $a > 0$. Consider a *uniformly random* path $\{u_0, u_1, \ldots, u_k, \ldots\}$ in $T_\infty$: $u_0$ is the root, and, for every $i \geq 0$, $u_{i+1}$ is a uniform random child of $u_i$. This implies that for all $k \geq 0$, $u_k$ is a uniform node in $\mathscr{L}_k$, the set of nodes $k$ levels away from the root in $T_\infty$. Let

$$L_{u_k} = \prod_{e \in \pi(u_k)} p_e = \prod_{e \in \pi(u_k)} e^{-E_e}. \tag{6.13}$$

By definition of $P_m(k,h)$, we have

$$\mathbf{E}P_m(k,h) \;=\; d^k \cdot \mathbf{P}\left\{\mathrm{Bin}(n, L_{u_k}) \geq m, D_{u_k} \geq h\right\}.$$

The randomness coming from the binomial random variables is irrelevant for the order of precision we are after. Indeed, for any $\xi_1 \in [0,1]$, we have

$$\mathbf{E}P_m(k,h) \;\leq\; d^k \cdot \mathbf{P}\left\{L_{u_k} \geq \xi_1, D_{u_k} \geq h\right\} + d^k \cdot \sup_{\xi \leq \xi_1} \mathbf{P}\left\{\mathrm{Bin}(n,\xi) \geq m\right\}. \tag{6.14}$$

In particular, if we set

$$\xi_1 = \frac{m d^{-k/m}}{e n^{1+a/m}}, \tag{6.15}$$

the second term of (6.14) is easily bounded as follows

$$\sup_{\xi \leq \xi_1} \mathbf{P}\left\{\mathrm{Bin}(n,\xi) \geq m\right\} \;\leq\; \mathbf{P}\left\{\mathrm{Bin}(n,\xi_1) \geq m\right\} \;\leq\; \binom{n}{m} {\xi_1}^m \;\leq\; \left(\frac{en\xi_1}{m}\right)^m \;=\; \frac{d^{-k}}{n^a}.$$

As a consequence, by definition of $D_{u_k}$ and (6.13),

$$\mathbf{E}P_m(k,h) \;\leq\; d^k \cdot \mathbf{P}\left\{\sum_{e \in \pi(u_k)} Z_e \geq h,\; \sum_{e \in \pi(u_k)} E_e \leq -\log \xi_1\right\} + \frac{1}{n^a}. \tag{6.16}$$

We shall now focus on the first term of (6.16). This kind of tail probability is treated by the theory of large deviations presented in Chapter 2. Recall that $\Delta_\phi(-a)$ is bounded by Lemma 6.2. So, there exists a constant $A > 0$ such that, for all $k \leq t \log n$ and $h \leq \alpha \log n$ with $(\alpha, t) \in \Delta_\phi(-a)$, we have, by (6.15),

$$-\log \xi_1 \;=\; \left(1 + \frac{a}{m}\right) \log n + 1 - \log m + \frac{k}{m} \log d \;\leq\; \log n \left(1 + \frac{A}{m}\right),$$

for $n$ large enough. Hence, rewriting (6.16), we have

$$\mathbf{E}P_m(k,h) \;\leq\; d^k \cdot \mathbf{P}\left\{\sum_{e \in \pi(u_k)} Z_e \geq h,\; \sum_{e \in \pi(u_k)} E_e \leq \log n \left(1 + \frac{A}{m}\right)\right\} + \frac{1}{n^a}. \tag{6.17}$$

By assumption, $\{E_e, e \in \pi(u_k)\}$ is a family of i.i.d. random variables. It is *not* the case for $\{Z_e, e \in \pi(u_k)\}$, and hence, not for $\{(Z_e, E_e), e \in \pi(u_k)\}$ either. However, by Lemma 6.1, the maximum number of nodes with less than $d$ children lying on a path down the root with $N \geq m(n)$ is $o(\log n)$ with probability at least $1 - n^{-\omega}$, for some $\omega \to \infty$ as $n \to \infty$. It follows that for i.i.d. random vectors $(Z_i^c, E_i)$, $i \geq 1$, distributed like $(Z^c, E)$, for $k \sim t \log n$ and $h \sim \alpha \log n$, for any $\delta > 0$, and $n$ large enough,

$$\mathbf{E}P_m(k,h) \leq d^k \cdot \mathbf{P}\left\{\sum_{i=1}^{k} Z_i^c \geq \left(\frac{\alpha}{t} - \delta\right) k, \sum_{i=1}^{k} E_i \leq \left(\frac{1}{t} + \delta\right) k\right\} + \frac{d^k}{n^\omega} + \frac{1}{n^a}. \quad (6.18)$$

Therefore, by Theorem 2.2, we have, for any $\delta > 0$, and $n$ large enough,

$$\mathbf{E}P_m(k,h) \ \leq \ \exp\left(k \log d - kI\left(\frac{\alpha}{t} - \delta, \frac{1}{t} + \delta\right)\right) + \frac{d^k}{n^\omega} + \frac{1}{n^a}.$$

Moreover, if

$$\left|\frac{h}{\log n} - \alpha\right| \xrightarrow[n\to\infty]{} 0 \quad \text{and} \quad \left|\frac{k}{\log n} - t\right| \xrightarrow[n\to\infty]{} 0$$

uniformly in a compact set, we can find $n$ large enough for any $\alpha$ and $t$ in the same set. Now, by definition, $\phi(\alpha, t) > -a$ on $\Delta_\phi(a)$ and hence $I(\cdot,\cdot)$ is finite and continuous at $(\frac{\alpha}{t}, \frac{1}{t})$, by Lemma 6.2 (see also Dembo and Zeitouni, 1998). Thus, since $\delta$ was arbitrary, for every $\epsilon > 0$, there exists $n$ large enough that

$$\mathbf{E}P_m(k,h) \ \leq \ \exp\left(k \log d - kI\left(\frac{\alpha}{t}, \frac{1}{t}\right) + \frac{\epsilon}{2}\right) \ = \ n^{\phi(\alpha,t)+\epsilon/2} + \frac{2}{n^a} \ \leq \ n^{\phi(\alpha,t)+\epsilon},$$

uniformly in any compact subset of $\Delta_\phi(-a) = \{(\alpha, t) : t > 0, \phi(\alpha, t) > -a\}$, where, as in (6.10),

$$\phi(\alpha, t) = t \log d - tI\left(\frac{\alpha}{t}, \frac{1}{t}\right).$$

This proves the lemma and the upper bound of Theorem 6.2. $\square$

We now focus on the lower bound and prove:

**Lemma 6.7.** *Let $m = m(n) \to \infty$ with $m = o(\log n)$. Let $k \sim t \log n$ and $h \sim \alpha \log n$ for some positive constants $t$ and $\alpha$. Let $\phi(\alpha, t)$ be given by (6.10). If $\phi(\alpha, t) > -\infty$, then $\mathbf{E}P_m(k,h) \geq n^{\phi(\alpha,t)+o(1)}$, as $n \to \infty$.*

We could handle the lower bound using techniques that are very similar to the ones we used in the case of weighted trees of Chapter 5. However, here, Lemma 6.1 permits to simplify the proof.

*Proof.* Recall that $u_k$ is a random node in $\mathscr{L}_k$, the set of nodes $k$ levels away from the root in $T_\infty$. We have

$$\mathbf{E}P_m(k,h) \;=\; d^k \cdot \mathbf{P}\left\{N_{u_k} \geq m, D_{u_k} \geq h\right\}.$$

Let $A_u$ be the event that all the nodes on the path from $u$ up to the root have $d$ children. Then, if $u \in \mathscr{L}_k$, $\mathbf{P}\left\{A_u \mid N_u\right\} \geq 1 - dk(1-p_d)^k$. Moreover, if $A_u$ occurs, all the weights on the path $\pi(u)$ are i.i.d. by construction, and we can use Cramér's theorem. We have

$$\begin{aligned}
\mathbf{E}P_m(k,h) \;&\geq\; d^k \cdot \mathbf{P}\left\{N_{u_k} \geq m, D_{u_k} \geq h, A_u\right\}\\
&=\; d^k \cdot \mathbf{P}\left\{N_{u_k} \geq m, D^c_{u_k} \geq h, A_u\right\},
\end{aligned}$$

where $D^c_u$ counts the depths using the variables $Z^c = Z^{(1,\dots,1)}$ of the embedding. Moreover,

$$\begin{aligned}
\mathbf{E}P_m(k,h) \;&\geq\; d^k \cdot \mathbf{P}\left\{N_{u_k} \geq m, D^c_{u_k} \geq h\right\} \cdot \mathbf{P}\left\{A_u \mid N_{u_k} \geq m, D^c_{u_k} \geq h\right\}\\
&=\; d^k \cdot \mathbf{P}\left\{N_{u_k} \geq m, D^c_{u_k} \geq h\right\} \cdot \mathbf{P}\left\{A_u \mid N_{u_k} \geq m\right\}\\
&\geq\; d^k \cdot \mathbf{P}\left\{N_{u_k} \geq m, D^c_{u_k} \geq h\right\} \cdot n^{o(1)}.
\end{aligned}$$

By definition $N_{u_k}$ is distributed as $\mathrm{Bin}(n, L_{u_k})$, where $L_{u_k}$ is defined in (6.13). As a consequence, for any $\xi_2$,

$$\mathbf{E}P_m(k,h) \;\geq\; d^k \cdot \mathbf{P}\left\{L_{u_k} \geq \xi_2, D^c_{u_k} \geq h\right\} \cdot \inf_{\xi \geq \xi_2} \mathbf{P}\left\{\mathrm{Bin}(n,\xi) \geq m\right\} \cdot n^{o(1)},$$

Choosing $\xi_2 = m/n$, we see that

$$\begin{aligned}
\inf_{\xi \geq \xi_2} \mathbf{P}\left\{\mathrm{Bin}(n,\xi) \geq m\right\} \;&=\; \mathbf{P}\left\{\mathrm{Bin}(n,\xi_2) \geq m\right\}\\
&\geq\; \mathbf{P}\left\{\mathrm{Bin}(n,\xi_2) \geq \mathbf{E}\mathrm{Bin}(n,\xi_2)\right\} \;=\; n^{o(1)},
\end{aligned}$$

and it follows that

$$\mathbf{E}P_m(k,h) \;\geq\; d^k \cdot \mathbf{P}\left\{L_{u_k} \geq \xi_2, D^c_{u_k} \geq h\right\} \cdot n^{o(1)}. \tag{6.19}$$

Therefore, it suffices to study $\mathbf{P}\left\{L_{u_k} \geq \xi_2, D^c_{u_k} \geq h\right\}$. Let $(Z^c_i, E_i)$, $i \geq 1$, be i.i.d. vectors distributed like $(Z^c, E)$. Then, by the definitions of $D^c_{u_k}$ and $L_{u_k}$,

$$\mathbf{P}\left\{L_{u_k} \geq \xi_2, D^c_{u_k} \geq h\right\} = \mathbf{P}\left\{\sum_{i=1}^k Z^c_i \geq h, \sum_{i=1}^k E_i \leq (1+o(1))\log n\right\},$$

since $-\log \xi_2 = (1+o(1))\log n$. By Cramér's theorem (Theorem 2.2) and (6.19), this yields,

$$\mathbf{E}P_m(k,h) \geq d^k \cdot \exp\left(-kI\left(\frac{\alpha}{t}, \frac{1-\epsilon}{t}\right) + o(k)\right) \cdot n^{o(1)},$$

for any $\epsilon > 0$ and $n$ large enough. Since $\epsilon$ is arbitrary and $I(\cdot,\cdot)$ is continuous where it is finite, the claim is proven:

$$\mathbf{E}P_m(k,h) \geq n^{\phi(\alpha,t)+o(1)},$$

where $\phi(\alpha, t)$ is given by (6.10). □

### 6.3.3 Logarithmic concentration: Proof of Theorem 6.3

The upper bound is straightforward using Markov's inequality and the uniform statement of Theorem 6.2. Indeed, by assumption, the convergence of $(h/\log n, k/\log n)$ is uniform in a compact set $\Gamma \subset \Delta_\phi(-a)$. For all $n$ large enough, and uniformly in the set $\Gamma$, $\mathbf{E}P_m(k,h) \leq n^{\phi(\alpha,t)+\epsilon/2}$, i.e.,

$$\sup_{(\alpha,t)\in\Gamma} \frac{\mathbf{E}P_m(k,h)}{n^{\phi(\alpha,t)}} \leq n^{-\epsilon/2}.$$

Hence

$$\mathbf{P}\left\{P_m(k,h) \geq n^{\phi(t,\alpha)+\epsilon}\right\} \leq \frac{\mathbf{E}P_m(k,h)}{n^{\phi(\alpha,t)+\epsilon}} \leq n^{-\epsilon/2}.$$

We now focus on the lower bound. We first prove a weaker version that will be boosted:

**Lemma 6.8.** *Let $\epsilon > 0$. Let $\alpha, t > 0$ such that $\phi(\alpha,t) > 0$. Then, for $k = \lfloor t\log n - 3t\log\log n\rfloor$, and $h \sim \alpha\log n$,*

$$\limsup_{n\to\infty} \mathbf{P}\left\{P_m(k,h) \leq n^{\phi(\alpha,t)-\epsilon}\right\} < 1.$$

*Proof.* In the previous section, one of the crucial arguments relies in the conditioning on the event $A_u$ that all the nodes along $\pi(u)$ are have $d$ children. Then, given $A_u$, we can use Cramér's theorem instead of the Gärtner–Ellis theorem. We use a similar argument, to relate $P_m(k,h)$ to a Galton–Watson process. By Lemma 6.1, we all the nodes such that $N_u \geq \log^2 n$ have degree $d$ with probability $1 - n^{\omega(n)}$. This is why we construct our Galton–Watson tree using the variables $(Z^c, E)$ of the embedding.

Let $B_u = -\log L_u = \prod_{e\in\pi(u)} p_e$. Let $\ell$ be a natural number to be chosen later. The individuals of the Galton–Waton process will be nodes of $\mathscr{L}_{s\ell}$, $s \geq 0$. A node $u$ is called *good* if either it is the root, or it lies $\ell$ levels below a good node $v$ and

$$D_u^c - D_v^c > \frac{\alpha\ell}{t} \qquad \text{and} \qquad B_u - B_v < \frac{\ell}{t}.$$

The set of good nodes is a Galton–Watson process. Let $G_s$ be the number of good nodes in the $s$-th generation, or at level $s\ell$ in $T_\infty$. Let $Y$ denote the progeny of an individual of the Galton–Watson process. Then, the expected progeny is

$$\mathbf{E}Y \;=\; d^\ell \mathbf{P}\left\{ \sum_{e\in\pi(u_\ell)} Z_e^c > \frac{\alpha\ell}{t}, \sum_{e\in\pi(u_\ell)} E_e < \frac{\ell}{t} \right\}$$

Hence, by Cramér's theorem (Theorem 2.2),

$$\mathbf{E}Y \;\geq\; d^\ell \cdot \exp\left(-\ell I\left(\frac{\alpha}{t},\frac{1}{t}\right) + o(\ell)\right) \;=\; \exp\left(\ell\log d - \ell I\left(\frac{\alpha}{t},\frac{1}{t}\right) + o(\ell)\right).$$

By assumption, $\phi(\alpha,t) > 0$ and $I(\alpha/t, 1/t) < \log d$. Then, for $\beta > 0$ small enough, there is $\ell$ large enough such that

$$\mathbf{E}Y \geq \exp\left(\ell\log d - \ell I\left(\frac{\alpha}{t},\frac{1}{t}\right) - \beta\ell\right) > 1. \tag{6.20}$$

Then, the process $\{G_s, s\geq 0\}$ of good nodes is supercritical.

Let $A$ be the event that all the nodes with $N_u \geq \log^2 n$ are of type $(1,\ldots,1)$. Let $F$ be the event that all the nodes with $nL_u \geq 2\log^2 n$ have $N_u \geq \log^2 n$. We have

$$\mathbf{P}\left\{P_m(k,h) \leq n^{\phi(\alpha,t)-\epsilon}\right\} \leq \mathbf{P}\left\{P_m(k,h) \leq n^{\phi(\alpha,t)-\epsilon}, A, F\right\} + \mathbf{P}\left\{\bar{A}\right\} + \mathbf{P}\left\{\bar{F}\right\}.$$

If both $A$ and $F$ occur, then, the nodes with $nL_u \geq 2\log^2 n$ all have $d$ children. Writing $r = r(n) = d^{-\ell} \cdot \log^2 n$,

$$\mathbf{P}\left\{P_m(k,h) \leq n^{\phi(\alpha,t)-\epsilon}, A, F\right\} \leq \mathbf{P}\left\{P_r(k,h) \leq n^{\phi(\alpha,t)-\epsilon}, A, F\right\}.$$

On the event $A$, all the variables influencing $P_r(k,h)$ are distributed as $(Z^c, E)$. Also, by definition of $k = \lfloor t\log n - 3\log\log n\rfloor$, for any good node $u$ at level $\ell\lfloor k/\ell\rfloor$, $nL_u \geq 2\log^2 n$ for $n$ large enough. Hence, if $k = 0 \bmod \ell$, $G_{\lfloor k/\ell\rfloor}$ is a lower bound on $P_r(k,\ell)$ if $F$ occurs. If $k \neq 0 \bmod \ell$, the subtree of every good node lying at level $\lfloor k/\ell\rfloor$ contains at level $k$ a node with $N_u \geq r$. Thus, since the weights are non-negative, for any $k \geq 0$, $G_{\lfloor k/\ell\rfloor}$ is a lower bound for $P_r(k,h)$ on $A \cap F$. As a consequence,

$$\begin{aligned}
\mathbf{P}\left\{P_m(k,h) \leq n^{\phi(\alpha,t)-\epsilon}, A, F\right\} &\leq \mathbf{P}\left\{P_r(k,h) \leq n^{\phi(\alpha,t)-\epsilon}, A, F\right\}\\
&\leq \mathbf{P}\left\{G_{\lfloor k/\ell\rfloor} \leq n^{\phi(\alpha,t)-\epsilon}, A, F\right\}\\
&\leq \mathbf{P}\left\{G_{\lfloor k/\ell\rfloor} \leq n^{\phi(\alpha,t)-\epsilon}\right\}.
\end{aligned}$$

Now, by Lemma 6.1, for $n$ large enough, $\mathbf{P}\left\{\bar{A}\right\} \leq n^{-\omega}$, for some $\omega \to \infty$, as $n \to \infty$. Also, by the union and Chernoff's bounds,

$$\mathbf{P}\left\{\bar{F}\right\} \leq d^k \cdot \mathbf{P}\left\{\mathrm{Bin}\left(n, \frac{2}{n}\log^2 n\right) \leq \log^2 n\right\} \leq d^k e^{-\frac{1}{8}\log^2 n} \leq e^{-\frac{1}{10}\log^2 n},$$

for $n$ large enough. It follows that

$$\mathbf{P}\left\{P_m(k,h) \leq n^{\phi(\alpha,t)-\epsilon}\right\} \leq \mathbf{P}\left\{G_{\lfloor k/\ell\rfloor} \leq n^{\phi(\alpha,t)-\epsilon}\right\} + o(1),$$

as $n \to \infty$. Therefore, proving the claim reduces to show that the first term is strictly lower than one. For this purpose, we take advantage of asymptotic properties of supercritical Galton–Watson properties.

By Doob's limit law (see Chapter 3), there exists a random variable $W$ such that

$$\frac{G_s}{\mathbf{E}G_s} \xrightarrow[s\to\infty]{} W \quad \text{almost surely.}$$

The equation above gives us a handle on $G_{\lfloor k/\ell\rfloor}$ via the limit distribution $W$. In particular, for any $\epsilon > 0$, we have

$$\mathbf{P}\left\{G_{\lfloor k/\ell\rfloor} \leq n^{\phi(\alpha,t)-\epsilon}\right\} \; = \; \mathbf{P}\left\{G_{\lfloor k/\ell\rfloor} \leq \mathbf{E}G_{\lfloor k/\ell\rfloor} \cdot n^{o(1)-\epsilon}\right\},$$

since $\mathbf{E}G_{\lfloor k/\ell \rfloor} \geq n^{\phi(t)+o(1)}$. As a consequence,

$$\mathbf{P}\left\{P_m(k,h) \leq n^{\phi(\alpha,t)-\epsilon}\right\} \;\leq\; \mathbf{P}\left\{\frac{G_{\lfloor k/\ell \rfloor}}{\mathbf{E}G_{\lfloor k/\ell \rfloor}} = o(1)\right\} \xrightarrow[k\to\infty]{} \mathbf{P}\left\{W=0\right\}.$$

The random variable $W$ is characterized by the Kesten–Stigum theorem (Theorem 3.4). In particular, writing $Y$ for the distribution of the progeny, $\mathbf{E}\left[Y \log(1+Y)\right] < \infty$ since $Y$ is bounded, and hence, $\mathbf{P}\left\{W=0\right\} = q$, the extinction probability of the coupled Galton–Watson process. Since the process is supercritical by (6.20), we have $q < 1$. □

We now intend to boost the bound given by Lemma 6.8. Consider $\mathscr{L}_\ell$, the set of nodes $\ell$ levels away from the root, for $\ell = \ell(n) = \lfloor \log\log n \rfloor$. Each one of $N_u$, $u \in \mathscr{L}_\ell$ is distributed as a binomial $\mathrm{Bin}(n, \xi_u)$ with $\xi_u \geq p_d^\ell$. Let $J_\ell$ be the good event that for each $u \in \mathscr{L}_\ell$, $N_u \geq n_\ell$, where $n_\ell = np_d^\ell/2$. Using the union bound, and then Chernoff's bound for binomial random variables (Chernoff, 1952; Janson et al., 2000), we see that

$$1 - \mathbf{P}\left\{J_\ell\right\} \;=\; \mathbf{P}\left\{\min_{u\in\mathscr{L}_\ell} N_u \leq n_\ell\right\} \;\leq\; d^\ell \cdot \mathbf{P}\left\{\mathrm{Bin}(n, p_d^\ell) \;\leq\; n_\ell\right\} \;\leq\; d^\ell \cdot e^{-n_\ell/2}.$$

Let $T_\infty(u)$ be the subtree of $T_\infty$ rooted at $u$. Given the values of the first $\ell$ symbols of each string, the subtrees $T_\infty(u)$, $u \in \mathscr{L}_\ell$ are independent. Moreover, conditioning on $J_\ell$, each one of these trees behave like a weighted trie with $N_u \geq n_\ell$ sequences. Let $P_{n,m}^u(k,h)$ be the number of nodes $v \in \mathscr{L}_k \cap T_\infty(u)$ such that $N_v \geq m$ and $D_v \geq h$. Since the weights are non-negative, we have

$$\begin{aligned}
\mathbf{P}\left\{P_{n,m}^u(k,h) \leq n^{\phi(t,\alpha)-\epsilon}\right\} \;&\leq\; \mathbf{P}\left\{P_{n_\ell,m}(k-\ell,h) \leq n^{\phi(\alpha,t)-\epsilon}\right\} \\
&\leq\; \mathbf{P}\left\{P_{n_\ell,m}(k-\ell,h) \leq n_\ell^{\phi(\alpha,t)-\epsilon+o(1)}\right\},
\end{aligned}$$

since $n/n_\ell = 2p_d^{\lfloor \log\log n \rfloor}$. Hence, for $n$ large enough, since $k - \ell \sim t\log n$ and $h \sim \alpha \log n$,

$$\mathbf{P}\left\{P_{n,m}^u(k,h) \leq n^{\phi(\alpha,t)-\epsilon}\right\} \;\leq\; \mathbf{P}\left\{P_{n_\ell,m}(k-\ell,h) \leq n_\ell^{\phi(t,\alpha)-2\epsilon}\right\} \;\leq\; q < 1,$$

by Lemma 6.8. However, if $P^u_{n,m}(k,h)$ is large for any of the nodes $u \in \mathscr{L}_\ell$, then $P_m(k,h)$ is large as well:

$$\mathbf{P}\left\{P_{n,m}(k,h) \le n^{\phi(t,\alpha)-\epsilon}\right\} \;\le\; \mathbf{P}\left\{\forall u \in \mathscr{L}_\ell, P^{(u)}_{n,m}(k,h) \le n^{\phi(t,\alpha)-\epsilon}\right\} \;\le\; q^{d^\ell} = o(1),$$

by independence. This finishes the proof of Theorem 6.3.

## 6.4 How long is a spaghetti?

### 6.4.1 Behavior and geometry

The behavior of the spaghettis is radically different from the one observed in the core. This is because the number of sequences each one of them stores is at most $m(n)$. There are two main questions of interest about the spaghettis. Of course, in preparation for the proof of Theorem 6.1, we shall study their maximum weighted height. But in order to acquire a deeper understanding of the situation, we will first look at the profile, not of a single trie, but of a forest of independent tries.

Let $T^1, T^2, \ldots, T^n$ be $n$ independent $b$-tries. We assume that $T^i$ is a weighted $b$-trie on $m_i = m_i(n)$ sequences generated by a memoryless source with distribution $\{p_1, \ldots, p_d\}$. Also, we assume that for all $i$, $m/d \le m_i \le m$. The roots of $T^i$, $1 \le i \le n$, all lie at level zero. Then, we let $P^s(k,h)$ count the number of nodes $u$ lying at level $k$ in any $T^i$ and such that $D_u \ge h$. Since $T^i$ is a $b$-trie, we only count the nodes for which $N_u \ge b+1$. For now, we are only interested in $\mathbf{E}P^s(k,h)$, when $k \sim \rho \log n$ and $h \sim \gamma \log n$. The behavior of the spaghettis is tightly related to that of $(b+1)$-tuples of strings. Recall that $Q(b+1) = \sum_{i=1}^d p_i^{b+1}$ is the probability that $b+1$ characters generated by the source $\{p_1, \ldots, p_d\}$ are identical. This is why the random variable of interest here is

$$Z^b = \begin{cases} Z^s_A & \text{w.p.} \quad Q(b+1) \\ -\infty & \text{otherwise,} \end{cases} \tag{6.21}$$

where $A \in \{1, \ldots, d\}$ is a random character generated by the memoryless source with probability distribution $\{p_1, \ldots, p_d\}$. Recall that the vector $\mathcal{Z}^s = (Z^s_1, \ldots, Z^s_d)$

is distributed as $(Z_1^{\sigma_1}, \ldots, Z_d^{\sigma_d})$ where $\sigma_i$ the permutation of $(1, 0, \ldots, 0) \in \{0,1\}^d$ with the one in the $i$-th position. Let $\Lambda_b^\star(\cdot)$ be the rate function associated with the variable $Z^b$, and recall that $\mathcal{D}^o_{\Lambda_b^\star}$ is the interior of the domain where it is finite.

**Theorem 6.4.** *Let $T^i$, $1 \le i \le n$, be a forest of $n$ independent tries. Let $T^i$ store $m_i = m_i(n)$ sequences. Assume that $m/d \le m_i \le m$ for all $1 \le i \le n$. Let $k \sim \rho \log n$ and $h \sim \gamma \log n$, as $n \to \infty$, for positive constants $\rho$ and $\gamma$. If $\gamma/\rho \in \mathcal{D}^o_{\Lambda_b^\star}$, then,*

$$\mathbf{E}P^s(k,h) = n^{\psi(\gamma,\rho)+o(1)},$$

*as $n \to \infty$, where $\psi(\gamma,\rho) = 1 - \rho I_b(\gamma/\rho)$, $\Lambda_b^\star(\cdot)$ is the rate function associated with the variable $Z^b$, and $I_b(x) = \inf\{\Lambda^\star(x') : x' > x\}$.*

The logarithmic profile of our forest of tries is shown of Figure 6.6. Observe in particular that the logarithmic profile decreases linearly along any fixed direction $\gamma/\rho$. In other words, the point $(0,0,1)$ casts a cone of projections on the horizontal plane. There is a *preferred* direction, corresponding to $(\gamma_b, \rho_b, 0)$ such that

$$\gamma_b = \sup_{\gamma,\rho>0} \{\gamma : \psi(\gamma,\rho) \ge 0\}.$$

This point is especially important since it characterizes the maximum weighted height of $T^1, \ldots, T^n$. Let $H^1, \ldots, H^n$ be the weighted heights of $T^1, \ldots, T^n$, respectively, and define

$$S_{n,b} = \max\{H^i : 1 \le i \le n\}.$$

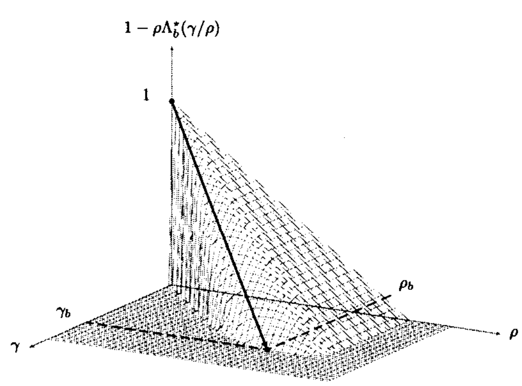


***Figure 6.6:*** *The profile generated by $n$ independent tries on roughly $m(n) = o(\log n)$ sequences each. We also show $\gamma_b$.*

**Theorem 6.5.** *Assume that $p_1 < 1$. Assume that $m(n) \to \infty$ and $m(n) = o(\log n)$. Let*

$$\gamma_b \stackrel{def}{=} \sup_{\gamma,\rho>0} \{\gamma : \psi(\gamma,\rho) \geq 0\} = \sup_{\gamma,\rho>0} \{\gamma : \rho\Lambda_b^\star(\gamma/\rho) \leq 1\}. \tag{6.22}$$

*Then, $S_{n,b} \sim \gamma_b \log n$ in probability, as $n \to \infty$. Furthermore, for every $\epsilon > 0$, there exists $\delta > 0$ such that, as $n \to \infty$,*

$$\mathbf{P}\left\{\frac{S_{n,b}}{\log n} \geq \gamma_b + \epsilon\right\} = O\left(n^{-\delta}\right). \tag{6.23}$$

The condition in the definition (6.22) of $\gamma_b$ reduces to finding the largest $\gamma$ such that there exists $\rho$ satisfying $\Lambda_b^\star(\rho) \leq \rho/\gamma$. In other words, if we plot $\rho \mapsto \Lambda^\star(\rho)$, then $1/\gamma_b$ is the slope of the most gentle line going through the origin and hitting the graph of $\Lambda^\star(\cdot)$, as shown on Figure 6.7. This yields the following characterization of $\gamma_b$:

**Lemma 6.9.** *Let $\Lambda_b^\star$ be the rate function associated with $Z^b$ defined in (6.21). Let $\gamma_b = \sup_{\gamma,\rho}\{\gamma : \rho\Lambda_b^\star(\gamma/\rho) \leq 1\}$. We have*

$$\begin{aligned} \gamma_b &= \sup\left\{\gamma : \exists\rho \quad \Lambda_b^\star(\rho) \leq \frac{\rho}{\gamma}\right\} \\ &= \sup\left\{\gamma : \exists\rho \quad I_b(\rho) \leq \frac{\rho}{\gamma}\right\} \\ &= \inf\left\{\gamma : \forall\rho \quad \Lambda_b^\star(\rho) > \frac{\rho}{\gamma}\right\}. \end{aligned}$$

*Proof.* The proof of the first inequality follows the lines of Lemma 4.2 and is not reproduced here. We have just proved the second equality. The third one follows from the min-max principle. □

Using this alternate definition of $\gamma_b$, we can characterize the value of $\gamma_b$.

**Lemma 6.10.** *Let $\Lambda_b^\star$ be the rate function associated with $Z^b$ defined in (6.21). Let $\gamma_b = \sup_{\gamma,\rho}\{\gamma : \rho\Lambda_b^\star(\gamma/\rho) \leq 1\}$. Assume that $Z^s$ is not almost surely null. If $Q(b) < 1$, then $\gamma_b \in (0,\infty)$. Otherwise, $\gamma_b = \infty$.*

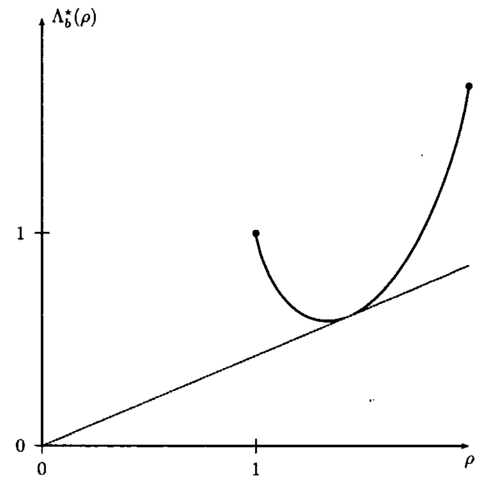


***Figure 6.7:*** *The constant $1/\gamma_b$ is the slope of the line going through the origin that is tangent to the curve $\{(\rho, \Lambda^\star(\rho))\}$.*

*Proof.* Recall from Chapter 2 that $\inf_\rho \Lambda_b^\star(\rho) = -\log \mathbf{P}\{Z^b > -\infty\}$. If $Q(b) < 1$, then $\inf_\rho \Lambda^\star(\rho) = -\log Q(b) > 0$. Moreover the infimum is reached at $\rho = \mathbf{E}\left[ Z^b \mid Z^b > -\infty \right] > 0$. The result follows (see Figure 6.7). On the other hand, if $Q(b) = 1$, then $\inf_\rho \Lambda^\star(\rho) = 0$ and $1/\gamma_b = 0$. □

### 6.4.2 The profile of a forest of tries: Proof of Theorem 6.4

In this section, we prove Theorem 6.4. We also define the notation that will be used in the proof of Theorem 6.5. The proof relies on the analysis of $(b+1)$-tuples of sequences. Let $\gamma, \rho > 0$ such that $\gamma/\rho \in \mathcal{D}^\circ_{\Lambda_b^\star}$. Let $k$ and $h$ be such that $h \sim \gamma \log n$, $k \sim \rho \log n$, as $n \to \infty$.

Consider any one of the $n$ tries. More particularly, consider a $(b+1)$-tuple of sequences generated by the source, $A^1, A^2, \ldots, A^{b+1}$, where $A^i = \{A^i_j, j \geq 0\}$ for $1 \leq i \leq b+1$. Let $u$ be a node at level $k$ in any of the tries. When taking a step one level down from $u$, we look at the next set $b+1$ of characters. Either they are identical, and the strings have followed the same path, or they are not and a $(b+1)$-tuple has been split. In the latter case, since we consider a $b$-trie, all the sequences are now stored in the nodes. In particular, this split $(b+1)$-tuple does not appear at levels deeper than $k$. Let $F_k$ be the event that all the characters in $k$-th position in the $(b+1)$-tuple are identical. Then, the $(b+1)$-tuple appears after level $k$ only if all $F_j$, $1 \le j \le k$ occur.

Now, taking one step down the trie also means that we have followed a weighted edge. The random weights are not i.i.d. on the path defined a string, since other $(b+1)$-tuples may interact and modify the type of the nodes. However, the influence of the interaction is negligible here. Indeed, as we have already seen, there are at most $m(n) = o(\log n)$ nodes in a trie (Lemma 6.1). As a consequence, in a single trie, the number of nodes whose type is not a permutation of $(1, 0, \dots, 0)$ is $o(\log n)$. In particular, since $Z$ is bounded, their influence is $o(\log n)$, which is negligible compared to $h = \gamma \log n$ for $\gamma > 0$ (see Lemma 6.1).

A node $u \in T^i$ lying $k$ levels away from the root is counted in $P^s(k, h)$ if $D_u \ge h$ and $u \in T^i$. This happens if there is at least one $(b+1)$-tuple stored in the subtree rooted at $u$, and the weighted depth is at least $h$. In other words, for all $i$, and a node $u$ at level $k$,

$$\begin{aligned}
\mathbf{P}\left\{D_u \ge h, u \in T^i\right\} &\le m^{b+1} \cdot \mathbf{P}\left\{\sum_{j=1}^{k} Z^j_{A^1_j} + o(\log n) \ge \gamma \log n, \bigcap_{1 \le j \le k} F_j\right\} \\
&= m^{b+1} \cdot \mathbf{P}\left\{\sum_{j=1}^{k} \left(Z^j_{A^1_j} - \infty \mathbf{1}[F_j^c]\right) + o(\log n) \ge \gamma \log n\right\}.
\end{aligned}$$

The summands on the right-hand side above are precisely distributed as $Z^b$ defined in (6.21). It follows that

$$\mathbf{P}\left\{D_u \ge h, u \in T^i\right\} \le m^{b+1} \cdot \mathbf{P}\left\{\sum_{j=1}^{k} Z^b_j + o(\log n) \ge \gamma \log n\right\}, \tag{6.24}$$

where $Z^b_j$, $1 \le j \le k$ are i.i.d. copies of $Z^b$. A lower bound is obtained by considering a single $(b+1)$-tuple for the trie $T^i$. This is possible since $m_i \ge m/d \to \infty$, and hence, for $n$ large enough, $m_i \ge b+1$ for all $i$. Then,

$$\mathbf{P}\left\{D_u \ge h, u \in T^i\right\} \ge \mathbf{P}\left\{\sum_{j=1}^{k} Z^b_j - o(\log n) \ge \gamma \log n\right\}. \tag{6.25}$$

Let $\delta > 0$ be arbitrary. There is $n$ large enough such that

$$\frac{\gamma \log n - o(\log n)}{k} \geq \frac{\gamma}{\rho} - \delta \qquad \text{and} \qquad \frac{\gamma \log n + o(\log n)}{k} \leq \frac{\gamma}{\rho} + \delta.$$

Use Cramér's Theorem in both (6.24) and (6.25), and observe that $m^{b+1} = n^{o(1)} = e^{o(k)}$ to obtain

$$e^{-kI_b(\gamma/\rho+\delta)+o(k)} \;\leq\; \mathbf{P}\left\{D_u \geq h, u \in T^i\right\} \;\leq\; e^{-kI_b(\gamma/\rho-\delta)+o(k)},$$

as $n \to \infty$. We have $n$ of these tries, and $I_b$ is continuous at $\gamma/\rho \in \mathcal{D}^o_{\Lambda_b^*}$. Hence,

$$\mathbf{E}P^s(k,h) = n^{1-\rho I_b(\gamma/\rho)+o(1)},$$

as $n \to \infty$. This completes the proof of Theorem 6.4.

### 6.4.3 The longest spaghetti: Proof of Theorem 6.5

THE UPPER BOUND. Let $\epsilon > 0$ be arbitrary, and write $\gamma' = \gamma_b + \epsilon$. We want to upper bound $S_{n,b}$. Observe that the tries $T^1, \ldots, T^n$ are not quite identically distributed. Indeed, their number of strings may vary slightly. We do not assume that the weighted height increases in the number of strings. By the union bound,

$$\mathbf{P}\left\{S_{n,b} \geq \gamma' \log n\right\} \leq nm \sup_{1 \leq i \leq n} \mathbf{P}\left\{H^i \geq \gamma' \log n\right\}. \tag{6.26}$$

Consider now any one of the $n$ tries. More particularly, consider a $(b+1)$-tuple of sequences generated by the source, $A^1, A^2, \ldots, A^{b+1}$, where $A^i = \{A^i_j, j \geq 0\}$ for $1 \leq i \leq b+1$. Let $W$ be the weighted height of the common path of this particular set of strings. By the same arguments we used in section 6.4.2:

$$W \leq \max_{\ell \geq 0} \left\{ \sum_{j=1}^{\ell} \left( Z^j_{A^1_j} - \infty \cdot \mathbf{1}[F^c_j] \right) \right\} + o(\log n), \tag{6.27}$$

where $Z^j$, $j \geq 1$, are i.i.d. copies of $Z^s$ defined in section 6.2. The summands in (6.27) are precisely i.i.d. copies of $Z^b$ defined by (6.21). Using the union bound, we see that for any $i$, and for $n$ large enough,

$$\mathbf{P}\left\{H^i > \gamma' \log n\right\} \;\leq\; m^{b+1} \cdot \mathbf{P}\left\{ \exists \ell : \sum_{j=1}^{\ell} Z^b_j \geq \gamma' \log n \right\},$$

where $\{Z_j^b, j \geq 1\}$ is a sequence of i.i.d. copies of $Z^b$. Note that the upper bound above is independent of $i$. Using (6.26) and the union bound once again, we obtain

$$\mathbf{P}\left\{S_{n,b} > \gamma' \log n\right\} \;\leq\; n \cdot m^{b+2} \cdot \sum_{\ell \geq 1} \mathbf{P}\left\{\sum_{j=1}^{\ell} Z_j^b \geq \gamma' \log n\right\},$$

Apply Cramér's theorem or Chernoff's bound (Theorem 2.2) to each one of the summands, and observe that $m^{b+2} = n^{o(1)}$:

$$\mathbf{P}\left\{S_{n,b} > \gamma' \log n\right\} \;\leq\; n^{1+o(1)} \cdot \sum_{\ell \geq 1} \exp\left(-\ell I_b\left(\frac{\gamma' \log n}{\ell}\right)\right). \tag{6.28}$$

We now split the sum on the right-hand side of (6.28) into two pieces, and then bound each one of them separately.

When $\ell$ is large, what prevents the sum to be large is the increasing probability that the path has been split. Recall that $\mathbf{P}\left\{Z^b > -\infty\right\} = Q(b+1)$, and hence $\inf_\rho I_b(\rho) = -\log Q(b+1)$. Let $\delta > 0$ to be fixed later. Let

$$K = K(n) = \frac{1+\delta}{-\log Q(b+1)} \cdot \log n.$$

We have

$$n^{1+o(1)} \sum_{\ell \geq K} \exp\left(-\ell I_b\left(\frac{\gamma' \log n}{\ell}\right)\right) = O\left(n^{1+o(1)} e^{\log Q(b+1) K}\right) = O\left(n^{-\delta/2}\right). \tag{6.29}$$

Now for the low values of $\ell$, we have to deal with the weights. Observe first that, by definition of $\gamma_b$, there exists $\beta > 0$ such that

$$\inf\left\{\frac{\ell}{\log n} \cdot I_b\left(\frac{\gamma'}{\ell / \log n}\right)\right\} \;\geq\; \inf_{\rho > 0}\left\{\rho \cdot I_b\left(\frac{\gamma'}{\rho}\right)\right\} \;=\; 1 + \beta.$$

We now choose $\delta$ small enough that $\delta < \beta/2$. Then, since $K = O(\log n)$,

$$n^{1+o(1)} \sum_{\ell \leq K} \exp\left(-\ell I_b\left(\frac{\gamma' \log n}{\ell}\right)\right) \;\leq\; K n^{-\beta/2} \;=\; O\left(n^{-\beta/4}\right). \tag{6.30}$$

Note that $\min\{\delta/2, \beta/4\} = \delta/2$. Plugging both (6.29) and (6.30) in (6.28) proves that

$$\mathbf{P}\left\{S_{n,b} > \gamma' \log n\right\} = O\left(n^{-\delta/2}\right) + O(n^{-\beta/4}) = O\left(n^{-\delta/2}\right),$$

which completes the proof of the upper bound (6.23).

THE LOWER BOUND. Let $\epsilon > 0$ and write $\gamma'' = \gamma_b - \epsilon$. By assumption, $m(n) \to \infty$, and hence, there exists $n$ large enough that $m(n)/d \geq b + 1$. We only consider one tuple from the each $b$-trie. We then have $n$ independent realizations of the random variable $W$ described in the previous section. We now have the lower bound

$$W \geq \max_{\ell \geq 0} \left\{ \sum_{j=1}^{\ell} Z^j_{A^1_j} - \infty \cdot \mathbf{1}[\exists k, k' : A^k_j \neq A^{k'}_j] \right\} - o(\log n), \tag{6.31}$$

where $Z^j$, $j \leq 1$ are i.i.d. copies of $Z^s$ defined in section 6.2. The largest of the $n$ independent copies of $W$ is a lower bound on $S_{n,b}$ (see remark next page). Let $X_i$, $1 \leq i \leq n$ denote the sequence of indicators that the $i$-th realization is at least $\gamma'' \log n$. Let $M = \sum_{i=1}^n X_i$. We intend to prove that $M \geq 1$ with probability tending to one as $n \to \infty$. For this purpose, we use the second moment method (see, e.g., Alon et al., 2000). We have,

$$\mathbf{E}M \;=\; n \cdot \mathbf{P}\left\{W \geq \gamma'' \log n\right\}.$$

Let $\{Z^b_j, j \geq 1\}$ be a sequence of i.i.d. copies of $Z^b$ defined by (6.21). Then,

$$\mathbf{E}M \;=\; n \cdot \mathbf{P}\left\{\exists \ell : \sum_{j=1}^{\ell} Z^b_j \geq \gamma'' \log n\right\} \geq n \cdot \mathbf{P}\left\{\sum_{j=1}^{\ell_0} Z^b_j \geq \gamma'' \log n\right\},$$

for any $\ell_0 \geq 1$. By the alternate definition of $\gamma_b$ provided by Lemma 6.9, there exists $\rho$ such that

$$\rho \cdot I_b\left(\frac{\gamma''}{\rho}\right) < 1.$$

In particular, if we set $\ell_0 = \lceil \rho \log n \rceil$, by Cramér's theorem,

$$\mathbf{E}M \geq n \cdot \mathbf{P}\left\{\sum_{j=1}^{\ell_0} Z^j_s \geq \gamma'' \log n\right\} = n \cdot \exp\left(-\ell_0 I_b\left(\frac{\gamma''}{\rho}\right) + o(\ell_0)\right) \xrightarrow[n\to\infty]{} \infty.$$

We now use the second moment method. By Chebychev's inequality,

$$\mathbf{P}\left\{M = 0\right\} \;\leq\; \mathbf{P}\left\{M - \mathbf{E}M \leq \mathbf{E}M\right\} \;\leq\; \frac{\mathbf{Var}\,[M]}{(\mathbf{E}\,[M])^2}.$$

However, $M$ is a sum of i.i.d. random variables, and we have $\mathbf{E}M = n\mathbf{E}X_1$ and $\mathbf{Var}[M] = n\mathbf{Var}[X_1]$. Also, $\mathbf{Var}[X_1] = \mathbf{E}X_1 - (\mathbf{E}X_1)^2$. It follows that

$$\mathbf{P}\{M=0\} \le \frac{1}{\mathbf{E}M} - \frac{1}{n} \xrightarrow[n\to\infty]{} 0.$$

As a consequence, with probability tending to one as $n \to \infty$, $M \ge 1$, which completes the proof of the lower bound.

**Remark.** Note that we have never used the fact that the weighted height of a trie is increasing in the number of sequences. In the lower bound, this is made possible since the random variable $W$ accounts only for the weights of edges tied to non-branching nodes, i.e., whose a type is a permutation of $(1,0,\ldots,0)$. Enforcing the fact that the weighted height be increasing in the number or strings would not affect any of our main applications (see section 6.6). However, for binary tries, where

$$\mathcal{Z}^\tau = \begin{cases} (1,1) & \text{if} \quad \tau = (1,1), \quad \text{and} \\ (2,2) & \text{if} \quad \tau \in \{(1,0),(0,1)\}, \end{cases}$$

the weighted height is not increasing. Yet, Theorems 6.1 and 6.5 still hold.

## 6.5 The height of weighted tries

### 6.5.1 Projecting the profile

Recall the definitions of the core and spaghettis. Let $m = m(n) \to \infty$ with $m = o(\log n)$. The core $\mathcal{C}$ of a $b$-trie $T_{n,b}$ is the set of nodes $u \in T_{n,b}$ such that $N_u \ge m$. When removing $\mathcal{C}$ from $T_n$, we obtain a forest of trees, the *spaghettis* (see Figure 6.1). Each one of these trees is rooted at a node $u \in \partial\mathcal{C}$, the external node-boundary of $\mathcal{C}$ in $T_{n,b}$. In other words, the nodes $u \in \partial\mathcal{C}$ are the children of some node $v$ in the core, but are not themselves in the core. Recall that $c_b = \sup\{\alpha + \gamma_b\phi(\alpha,t)\}$, where

$$\gamma_b = \sup\{\gamma : \rho\Lambda_b^\star(\gamma/\rho) \le 1, \gamma > 0, \rho > 0\}.$$

Alternate definitions of $\gamma_b$ are given in Lemma 6.9.

Theorem 6.1 can be interpreted as follows. Consider a point $(\alpha, t, \phi(\alpha, t))$. This point is mapped on the horizontal plane going through the origin via a *projection.* The direction of the projection is given by the vector $(1, 0, -1/\gamma_b)$. The direction along the $t$-axis is actually irrelevant, and any direction $(1, x, -1/\gamma_b)$ gives the same $\alpha$-coordinate for the image of $(\alpha, t, \phi(\alpha, t))$. The constant $c_b$ is then largest $\alpha$-coordinate of these projections.

The projection is not a mere interpretation of the *formula* for $c_b$. Indeed, Theorem 6.4 shows that a set of $P_m(k, h)$ tries on about $m(n)$ sequences each has a logarithmic profile that decays linearly in every direction. We can also note that the actual profile induces a *preferred* direction of projection $(1, -1/\rho_b, -1/\gamma_b)$, as shown of Figure 6.6. The projection of points $(\alpha, t, \phi(\alpha, t))$ using this preferred direction is depicted in Figure 6.8.

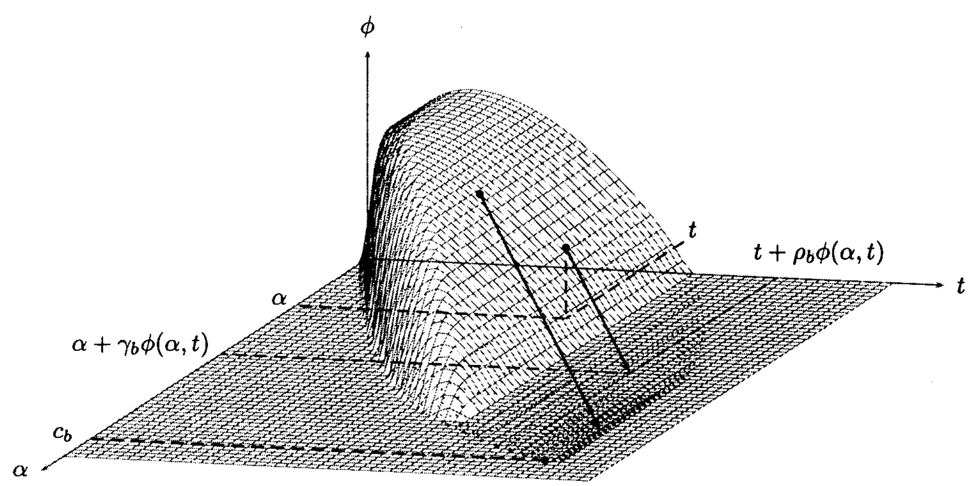


***Figure 6.8:*** *A geometric interpretation for the height: each point $(\alpha, t, \phi(\alpha, t))$ of the logarithmic profile of the core throws a line whose direction is given by $(1, -1/\rho_b, -1/\gamma_b)$. The line intersects the plane $\phi = 0$ at $(\alpha + \gamma_b\phi(\alpha, t)), t + \rho_b\phi(\alpha, t), 0)$. The constant $c_b$ is the largest coordinate of one of these point along the $\alpha$-axis.*

The definition of $c_b$ above follows from the proof. For the applications, it is useful to simplify it slightly.

**Proposition 6.1.** *We have* $c_b = \sup\{\alpha + \gamma_b \cdot \phi_o(\alpha, t)\}$, *where* $\phi_o(\cdot, \cdot)$ *is defined by*

$$\phi_o(\alpha, t) = t \log d - \Lambda^\star\left(\frac{\alpha}{t}, \frac{1}{t}\right).$$

*Proof.* Clearly, since $\phi(\cdot, \cdot)$ is concave, only the points $(\alpha, t)$ for which $\phi(\alpha, t) \geq 0$ matter. By definition,

$$I\left(\frac{\alpha}{t}, \frac{1}{t}\right) = \inf\left\{\Lambda^\star(x, y) : x > \frac{\alpha}{t}, y < \frac{1}{t}\right\}.$$

The function $\Lambda^\star(\cdot, \cdot)$ is continuous at every point $(\alpha/t, 1/t)$ such that $\phi(\alpha, t) \geq 0$. Hence for every such a point, $\phi_o(\alpha, t) \leq \phi(\alpha, t)$. As a consequence

$$c_b = \sup\{\alpha + \gamma_b\phi(\alpha, t)\} \geq \sup\{\alpha + \gamma_b\phi_o(\alpha, t)\}. \tag{6.32}$$

To devise the other inequality, we need only consider the points for which $\phi(\alpha, t) \neq \phi_o(\alpha, t)$, that is $(\alpha, t)$ such that $t\mathbf{E}E \leq 1$ and $\alpha \leq t\mathbf{E}Z$. It turns out that these points do not matter in the supremum since the value they account for is always dominated by some other one. Observe that $\mathbf{E}E \geq \mathcal{E}$, and by Lemma 6.2, $\phi(\cdot, \cdot)$ is concave and maximum at $(\alpha, 1/\mathcal{E})$, $\alpha \leq \mathcal{E}\mathbf{E}Z$. Therefore, for all $\alpha$ and $t$ such that $t\mathbf{E}E \leq 1$ and $\alpha \leq t\mathbf{E}Z$,

$$\alpha + \gamma_b\phi(\alpha, t) \leq \frac{\mathbf{E}Z}{\mathcal{E}} + \gamma_b\phi\left(\frac{\mathbf{E}Z}{\mathcal{E}}, \frac{1}{\mathcal{E}}\right).$$

It follows that the points for which $\phi(\alpha, t)$ and $\phi_o(\alpha, t)$ differ are irrelevant. Each single one of the relevant points in $\sup\{\alpha + \gamma_b\phi(\alpha, t)\}$ is also present in $\sup\{\alpha + \gamma_b\phi_o(\alpha, t)\}$. Thus, $c_b \leq \sup\{\alpha + \gamma_b\phi_o(\alpha, t)\}$. With (6.32), this completes the proof. □

### 6.5.2 Proof of Theorem 6.1

Put together, Lemmas 6.11 and 6.12 prove Theorem 6.1.

**Lemma 6.11.** *Let* $T_{n,b}$ *be a b-trie as defined in section 6.2. Let* $H_{n,b}$ *be its height. Then, for any* $\epsilon > 0$, $\mathbf{P}\{H_{n,b} \geq (c_b + \epsilon)\log n\} \to 0$ *as* $n \to \infty$.

*Proof.* Let $\epsilon > 0$ and write $c' = c_b + \epsilon$. Let $W_u$ denote the weighted height of the subtree rooted at $u$. Recall that $\mathcal{C}$ denotes the core. We have

$$\mathbf{P}\{H_{n,b} \geq c' \log n\} \;\leq\; \mathbf{P}\{\exists u \in \mathcal{C} : D_u + W_u \geq c' \log n\}.$$

Let $\mathcal{C}_k = \mathcal{C} \cap \mathscr{L}_k$, where $\mathscr{L}_k$ is the set of nodes $k$ levels away from the root in $T_\infty$. Then,

$$\mathbf{P}\{H_{n,b} \geq c' \log n\} \leq \mathbf{P}\{\exists k, u \in \mathcal{C}_k : D_u + W_u \geq c' \log n\}.$$

We can immediately restrict the range of $k$. Indeed, when $k$ is too large, it is unlikely that there is even one node $u$ in the $\mathcal{C}_k$. By Lemma 6.2, $\{(\alpha, t) : \phi(\alpha, t) \geq -\epsilon\}$ is contained in a bounded set. Pick $t$ large enough that $\phi(0, t) \leq -\epsilon < 0$. Let $K = K(n) = \lceil t \log n \rceil$. Then,

$$\begin{aligned}
\mathbf{P}\{\exists k \geq K, u \in \mathcal{C}_k : D_u + W_u \geq c' \log n\} &\leq \mathbf{P}\{|\mathcal{C}_K| > 0\} \\
&\leq \mathbf{E} P_m(0, K) \\
&\leq n^{-\epsilon + o(1)},
\end{aligned}$$

by Theorem 6.2. Let $\mathcal{C}_k(h) = \{u \in \mathcal{C}_k : D_u \geq h\}$. By the union bound,

$$\begin{aligned}
\mathbf{P}\{H_{n,b} \geq c' \log n\} &\leq \sum_{k \leq K} \mathbf{P}\{\exists u \in \mathcal{C}_k : D_u + W_u \geq c' \log n\} + o(1) \qquad (6.33) \\
&= \sum_{k \leq K} \underbrace{\mathbf{P}\left\{\exists h : \mathcal{C}_k(h) \neq \emptyset, h + \max_{u \in \mathcal{C}_k(h)} W_u \geq c' \log n\right\}}_{R(k)} + o(1).
\end{aligned}$$

Let $k \leq K$, and consider the corresponding term $R(k)$ in the sum above. Let $\mathcal{F}_k$ be the $\sigma$-algebra generated by the first $k$ symbols of the $n$ strings. Then,

$$R(k) = \mathbf{E}\left[\mathbf{P}\left\{\exists h : \mathcal{C}_k(h) \neq \emptyset, h + \max_{u \in \mathcal{C}_k(h)} W_u \geq c' \log n \;\middle|\; \mathcal{F}_k\right\}\right].$$

However, given $\mathcal{F}_k$, $\max\{W_u : u \in \mathcal{C}_k(h)\}$ is distributed like the longest of $P_m(k, h)$ independent weighted tries, each on at most $m(n)$ sequences. Therefore, by Theorem 6.5, for any $\beta > 0$ there exists $\delta > 0$ such that

$$\mathbf{P}\left\{\max_{u \in \mathcal{C}_k(h)} W_u \geq (\gamma_b + \beta) \log P_m(k, h) \;\middle|\; \mathcal{F}_k\right\} \leq e^{-\delta \log P_m(k,h)},$$

where $\gamma_b$ defined by (6.22). This bound is weak when $P_m(k, h)$ is small. In such a case, we shall rather use

$$\mathbf{P}\left\{\max_{u \in \mathcal{C}_k(h)} W_u \geq \frac{\epsilon}{2} \log n \;\middle|\; \mathcal{F}_k\right\} \leq n^{-\delta\epsilon/(4\gamma_b + 4\beta)}$$

when $P_m(k,h) \leq n^{\epsilon/4}$. It follows that, in any case, and for $\beta \leq \gamma_b$,

$$R(k) \leq \mathbf{P}\left\{\exists h : h + (\gamma_b + \beta)\log P_m(k,h) \geq \left(c_b + \frac{\epsilon}{2}\right)\log n\right\} + n^{-\delta\epsilon/(4\gamma_b)}. \tag{6.34}$$

We now bound the first term of (6.34). The full range for $k$ and $h$ is obtained by setting $k = \lfloor t \log n \rfloor$ and $h = \alpha \log n$, and letting $t$ and $\alpha$ vary. By definition of $c_b = \sup\{\alpha + \gamma_b \phi(\alpha,t)\}$, we have for all $h = \alpha \log n$ and $k = \lfloor t \log n \rfloor$,

$$\mathbf{P}\left\{\frac{\log P_m(k,h)}{\log n} \geq \frac{1}{\gamma_b + \beta}(c_b - \alpha) + \frac{\epsilon/2}{\gamma_b + \beta}\right\} \leq \mathbf{P}\left\{\frac{\log P_m(k,h)}{\log n} \geq \frac{\gamma_b \phi(\alpha,t) + \epsilon/2}{\gamma_b + \beta}\right\}.$$

The weights are bounded and there exists $A$ such that for all $k \leq t_o \log n$, $h \leq A \log n$. A bound that is uniform on compact sets is given by Theorem 6.3, and for $\beta > 0$ small enough,

$$\sup_{t \leq t_o, h \leq A} \mathbf{P}\left\{\frac{\log P_m(k,h)}{\log n} \geq \frac{1}{\gamma_b + \beta}(c_b - \alpha) + \frac{\epsilon/2}{\gamma_b + \beta}\right\} \leq n^{\epsilon/(4\gamma_b)+o(1)}.$$

As a consequence, recalling (6.33) and (6.34),

$$\begin{aligned}
\mathbf{P}\{H_{n,b} \geq c' \log n\} &\leq \sum_{k \leq t_o \log n, h \leq A \log n} n^{-\epsilon/(4\gamma_b)+o(1)} + \sum_{k \leq t_o \log n, h \leq A \log n} n^{-\delta\epsilon/(4\gamma_b)} \\
&\leq O\left(n^{-\epsilon/(5\gamma_b)} \log^2 n\right) + O\left(n^{-\delta\epsilon/(4\gamma_b)} \log^2 n\right).
\end{aligned}$$

This completes the proof of the upper bound. □

**Lemma 6.12.** *Let $T_{n,b}$ be a b-trie as defined in section 6.2. Let $H_{n,b}$ be its height. Then, for any $\epsilon > 0$, $\mathbf{P}\{H_{n,b} \leq (c_b - \epsilon)\log n\} \to 0$ as $n \to \infty$.*

*Proof.* Let $\epsilon > 0$. Recall that, by definition, $c_b = \sup\{\alpha + \gamma_b \cdot \phi(\alpha,t) : t, \alpha > 0\}$. Therefore, there exists $(\alpha_o, t_o)$ such that

$$\alpha_o + \gamma_b \cdot \phi(\alpha_o, t_o) \geq c_b - \epsilon/2.$$

Let $\alpha_o$ and $t_o$ now be fixed. Let $k = \lfloor t_o \log n - 3t_o \log\log n \rfloor$ (in order to use Theorem 6.3) and $h = \alpha_o \log n$. Let $\mathcal{F}_k$ be the $\sigma$-algebra generated by the first $k$ characters of the $n$ strings. Consider the $N' = P_m(k,h)$ nodes $u$ at level $k$ for which $N_u \geq m$,

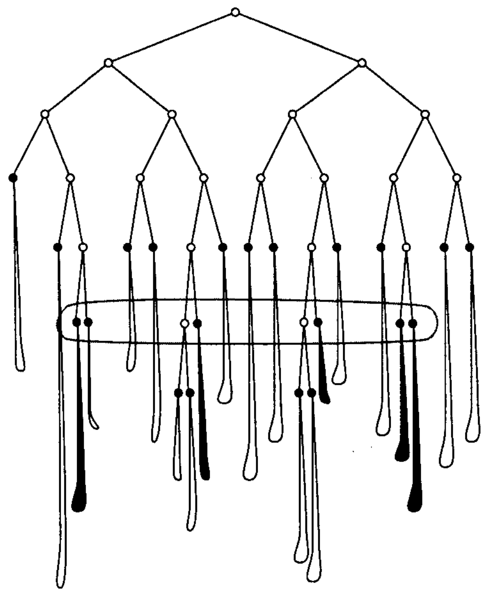

***Figure 6.9:*** *The structure of the lower bound: find some level k such that the trees starting at $\mathscr{L}_k$ are tall enough. Only those (colored) are considered, the others are ignored.*

$D_u \geq h$. Conditioning on $\mathcal{F}_k$, the $P_m(k,h)$ trees rooted at these nodes are independent. Let $S_{N',b}$ be the weighted height of the tallest of these trees. We want to show that $h + S_{N',b}$ is a good enough lower bound on $H_{n,b}$.

So it suffices to lower bound $S_{N',b}$. We are in the situation we have studied in Section 6.4, and we intend to apply Theorem 6.5. Let $\delta > 0$ and $n' = n^{\phi(\alpha_o,t_o)-\delta}$. The idea of the lower bound is pictured in Figure 6.9. We have

$$\begin{aligned} \mathbf{P}\left\{ \frac{S_{N',b}}{\log n'} \leq \gamma_b - \delta \;\middle|\; \mathcal{F}_k \right\} \;\leq\; & \mathbf{P}\left\{ \frac{S_{N',b}}{\log n'} \leq \gamma_b - \delta \;\middle|\; \mathcal{F}_k, P_m(k,h) \geq n^{\phi(\alpha_o,t_o)-\delta} \right\} \\ & + \mathbf{1}[P_m(k,h) \leq n^{\phi(\alpha_o,t_o)-\delta}]. \end{aligned}$$

Taking expected values, we obtain

$$\mathbf{P}\left\{ \frac{S_{N',b}}{\log n'} \leq \gamma_b - \delta \right\} \leq \mathbf{P}\left\{ \frac{S_{n',b}}{\log n'} \leq \gamma_b - \delta \right\} + \mathbf{P}\left\{ P_m(k,h) \leq n^{\phi(\alpha_o,t_o)-\delta} \right\}. \quad (6.35)$$

It only remains to bound both terms on the right-hand side of the equation (6.35). By Theorem 6.3 and the definition of $k$ and $h$, $\mathbf{P}\left\{ P_m(k,h) \leq n^{\phi(t,\alpha)-\delta} \right\} = o(1)$. Also, by Theorem 6.5,

$$\mathbf{P}\left\{ \frac{S_{n',b}}{\log n'} \leq \gamma_b - \delta \right\} \xrightarrow[n\to\infty]{} 0.$$

Therefore, with probability $1 - o(1)$,

$$S_{N',b} \;\geq\; (\gamma_b - \delta)\log n' \;=\; (\gamma_b - \delta) \cdot (\phi(\alpha_o, t_o) - \delta) \cdot \log n.$$

We have not specified $\delta$ yet, and we pick $\delta$ small enough that

$$(\gamma_b - \delta) \cdot (\phi(\alpha_o, t_o) - \delta) > \gamma_b \phi(\alpha_o, t_o) - \epsilon/2.$$

The weighted height of $T_{n,b}$ is at least $h + S_{N',b}$. It follows that, with probability $1 - o(1)$,

$$\frac{H_{n,b}}{\log n} \geq \alpha_o + (\gamma_b - \delta) \cdot (\phi(\alpha_o, t_o) - \delta) \geq c_b - \epsilon,$$

by our choice of $\delta$. This completes the proof of the lower bound. □

## 6.6 Applications

### 6.6.1 Standard $b$-tries

We shall first consider simple well-known examples. We start with the case of standard, i.e. unweighted, tries. We show that the following theorem follows from Theorem 6.1.

**Theorem 6.6.** *Consider an unweighted b-trie $T_{n,b}$ on $n$ independent sequences of characters of $\{1, \dots, d\}$ generated by a memoryless source with distribution $p_1 \geq \cdots \geq p_d > 0$. Let $H_{n,b}$ denote the height of $T_{n,b}$. Then,*

$$\frac{H_{n,b}}{\log n} \xrightarrow[n\to\infty]{} \frac{b+1}{-\log Q(b+1)}$$

*in probability, as $n \to \infty$.*

Theorem 6.6 is due to Szpankowski (1991). The case $b = 1$ was proved by Pittel (1985). See also Devroye et al. (1992).

**Remark.** Theorem 6.6 has first been proved by considering the longest prefix of $(b+1)$-tuples of sequences, which is exactly what we do for the analysis of the spaghettis. It is interesting to note that for this case, one can obtain tight bounds on the height without distinguishing the core from the spaghettis. One of the reasons is that the weights are identical for all the edges.

*Proof.* Here, we assume that $Z = 1$ almost surely. Then, $\phi(\alpha, t)$ is just the logarithmic profile studied by Park et al. (2006) in the binary case, or Broutin and Devroye (2007a).

THE CORE OF THE TRIE. We can compute the generating function of the cumulants: for any $\lambda, \mu \in \mathbb{R}$,

$$\Lambda(\lambda, \mu) = \log \mathbf{E}\left[e^{\lambda Z + \mu E}\right] = \lambda + \log \sum_{i=1}^{d} p_i^{-\mu} - \log d.$$

Then, the associated convex dual $\Lambda^\star$ is given by

$$\Lambda^\star(x, y) = \sup_{\lambda,\mu} \left\{ \lambda(x-1) + \mu y - \log \sum_{i=1}^{d} p_i^{-\mu} \right\} + \log d.$$

It follows that $\Lambda^\star(x, y)$ is infinite unless $x = 1$. Writing $\mu = \mu(y)$ for the unique solution of

$$y = \frac{\sum_{i=1}^{d} \log p_i \cdot p_i^{-\mu}}{\sum_{i=1}^{d} p_i^{-\mu}},$$

we have

$$\Lambda^\star(1, y) = \mu y - \log \sum_{i=1}^{d} p_i^{-\mu} + \log d.$$

By Proposition 6.1, it suffices to study $\phi_o$ instead of the more complicated $\phi$. By definition,

$$\phi_o(\alpha, \alpha) \;=\; \alpha \log d - \alpha \Lambda^\star\left(1, \frac{1}{\alpha}\right) \;=\; \mu(1/\alpha) + \alpha \log \sum_{i=1}^{d} p_i^{-\mu(1/\alpha)}. \qquad (6.36)$$

THE BEHAVIOR OF SPAGHETTIS. In an unweighted trie, we have $Z^s = 1$, and $Z^b = 1$ almost surely. Therefore, for all $\lambda$,

$$\Lambda_b(\lambda) = \log \mathbf{E}\left[e^{\lambda}\right] + \log Q(b+1),$$

and hence $\Lambda_b^\star(x)$ is infinite unless $x = 1$, in which case, we have $\Lambda_b^\star(1) = -\log Q(b+1)$. Then, Lemma 6.9, we clearly have

$$\gamma_b = \sup\left\{\gamma : \exists \rho \quad \Lambda_b^\star(\rho) \le \frac{\rho}{\gamma}\right\} = \frac{1}{-\log Q(b+1)}.$$

THE OVERALL CONTRIBUTION. Now by Theorem 6.1, the height $H_{n,b}$ of a random $b$-trie is asymptotic to $c_b \log n$ in probability, where

$$c_b = \sup_{\alpha>0} \left\{ \alpha + \frac{\phi(\alpha,\alpha)}{-\log Q(b+1)} \right\}.$$

This reduces to finding $\alpha_o$ such that

$$\left. \frac{\partial\phi(\alpha,\alpha)}{\partial\alpha} \right|_{\alpha=\alpha_o} = \log Q(b+1).$$

By Lemma 6.4, $\alpha_o = Q(b+1)/\mathcal{E}(b+1)$, where

$$\mathcal{E}(b+1) = \sum_{i=1}^{d} p_i^{b+1} \log p_i.$$

Lemma 6.4 also implies that

$$c_b = \frac{b+1}{-\log Q(b+1)}.$$

This completes the proof of Theorem 6.6. For an illustration of this case, see Figure 6.3. □

**Example: symmetric $b$-tries.** When $p_1 = p_2 = \cdots = p_d = 1/d$, the functions $\Lambda^\star(\cdot,\cdot)$ and $\phi_o(\cdot,\cdot)$ are degenerate. Our framework works in this case. In particular, $\phi_o(\alpha,t)$ is degenerate: $\phi_o(\alpha,t) = -\infty$, unless $\alpha = 1/\log d$ and $t = 1/\log d$, where $\phi_o(\alpha,t) = 1$. In this case, $\log Q(b+1) = -b\log d$. It follows that

$$H_{n,b} \sim \left( \frac{1}{\log d} + \frac{1}{-\log Q(b+1)} \right) \log n = \left(1 + \frac{1}{b}\right) \log_d n$$

in probability, as $n \to \infty$. In such a case, the contribution of the spaghettis is $1/b$ times that of the core. For instance, with ordinary tries, $b = 1$ and the contribution of spaghettis is equivalent to that of the core. This result was first obtained by Régnier (1981) in the case of a Poisson number of sequences. Flajolet and Steyaert (1982) and Flajolet (1983) obtained the limit distribution. See also Devroye (1984) and Pittel (1985).

| | b | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 10 | 50 | 100 |
| $c_b(2)$ | 2.88539... | 2.16404... | 1.92359... | 1.58696... | 1.47154... | 1.45712... |
| $c_b(3)$ | 1.82047... | 1.36535... | 1.21365... | 1.00126... | 0.92844... | 0.91934... |
| $c_b(10)$ | 0.86858... | 0.65144... | 0.57905... | 0.47772... | 0.44298... | 0.43863... |

***Table 6.1:*** *Some numerical values for $c_b = c_b(d)$ the height of symmetric ordinary tries, as $b$ varies and $d \in \{1, 2, 10\}$.*

### 6.6.2 Efficient implementations of tries

The usual implementation of a trie uses an array for the branching structure of a node (Fredkin, 1960). Although this always ensures $O(1)$ shunting of the strings, the space required may become an issue for large alphabets: many pointers would be left unused. To avoid this, one solution is to replace the array by variable size structures. The oldest solution due to de la Briandais (1959) uses a linked-list, and we shall call the implementation a *list-trie.* More recently, a second elegant solution has been proposed by Bentley and Sedgewick (1997), which uses binary search trees. It is known as the *bst-trie*, ternary search trie or TST for short.

Both structures aim at a trade-off between the storage space and the speed. In particular, the access time to children is no longer constant, and may even not be $O(1)$ when the alphabet is infinite. In this sense, list-tries and the TST may be seen as *high-level* tries whose edges are weighted to reflect the internal *low-level* structure of a node (see Figure 6.10). This point of view has been taken by Clément, Flajolet, and Vallée (1998, 2001) who analyzed thoroughly these *hybrid* implementations of tries under various models of randomness (see also Clément, 2000). In particular, they analyzed the average size and average depth. The question of the height of hybrid-tries was left open. We show that the heights of both the list-trie and TST follow from Theorem 6.1.

Let $\mathcal{A} = \{1, \dots, d\}$ be the alphabet. Let $\{A^i, 1 \le i \le n\}$ be the $n$ strings. In ordinary tries, that is, with the array implementation, the order of the sequences is irrelevant. This is not the case any more in either the list-trie or the TST. In the following, we distinguish the *nodes* that constitute the high-level trie structure from the *slots* which make the low-level structure of a node, whether this latter be a linked-list or a binary search tree.

We now describe the way the internal structure of a node is constructed, in both list-tries and TSTs. Consider a node $u$. The subtree rooted at $u$ stores a subset of the sequences $A^i$, $1 \leq i \leq n$. Let $\mathcal{N}_u \subset \{1, \ldots, n\}$ be the set of their indices. So, in particular, the cardinality of $u$ is $N_u = |\mathcal{N}_u|$. The internal structure of the node is built using the sequences in increasing order of their index (see Figure 6.10). For a node $u$ at level $k$ in $T_\infty$, only the $k$-th characters of each sequence are used. Only the distinct characters matter. Let $\mathcal{A}_u \subset \mathcal{A}$ be the set of distinct characters appearing at the $k$-th position in the sequences $A^i$, $i \in \mathcal{N}_u$. The characters in $\mathcal{A}_u$ are ordered by first appearance, and this induces a permutation $\sigma_u$ of $\mathcal{A}_u$. The internal structure of the node $u$ is built by successive insertions of the elements of $\sigma_u$ into an originally empty linked list, or binary search tree.

Both the list-tries and ternary search trees are built using the process we have just described. We shall now study each one of them more precisely.

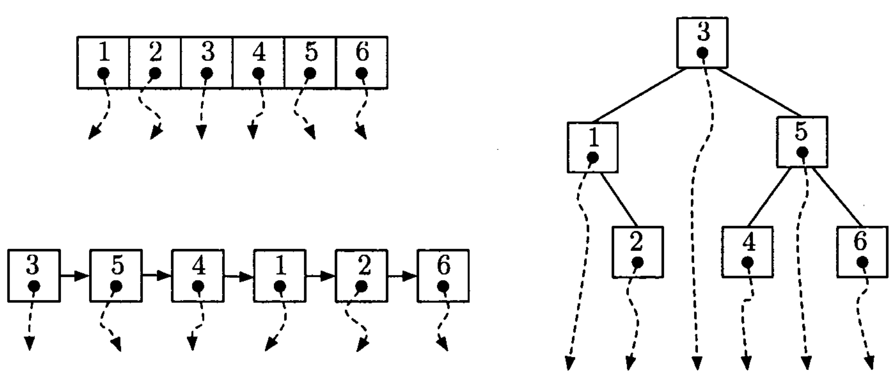


***Figure 6.10:*** *The different node structures used for the standard (top-left), list (bottom-left) and bst-trie (right) when the order of appearance of the characters is* 3, 5, 4, 1, 2 *and* 6. *The dashed arrows represent the pointers to further levels of the trie.*

### 6.6.3 List-tries

In the list-trie of de la Briandais (1959), the cost of branching to a character $a$ is just the index of $a$ is the permutation $\sigma_u$. For every node $u$, for which $\mathcal{A}_u = \mathcal{A}$, $\sigma_u$ is distributed as the sequence (in order) of first appearance of characters in an infinite string generated by the source. This fully describes the distribution of $Z$. That is, we have $Z_i$ is the index of $i$ in $\sigma$, and $Z = Z_K$, where $K$ is uniform in $\{1, \ldots, d\}$. Observe that when $|\mathcal{A}_u| = 1$, we have $Z = 1$.

**Theorem 6.7.** *Let $H_{n,b}$ be the weighted height of a list-trie on $n$ sequences. Let $Z$ be as described above. Then, $H_{n,b} \sim c_b \log n$ in probability, as $n \to \infty$, where*

$$c_b = \sup_{\alpha,t>0} \left\{ \alpha + \frac{\phi(\alpha, t)}{-\log Q(b+1)} \right\},$$

*and $\phi(\cdot,\cdot)$ is the logarithmic profile of the trie weighted with $Z$.*

The theorem explains and characterizes the first term of the asymptotic expansion of the height for all distributions $p_1, \ldots, p_d$ for $d < \infty$. For general distributions, it seems difficult to obtain a closed form for the height. However, $Z$ is a non uniform random draw of an element of $\{1, \ldots, d\}$, and maybe there is an other way to see the random variable that would lead to the a better description of $\phi(\alpha, t)$ and $c_b$. We shall obtain more concrete results for specific example.

**Example: symmetric list-tries.** In this case, for all $i$, we have $p_i = 1/d$ and $Z_i$ is uniform in $\{1, \ldots, d\}$. Therefore, for any $\lambda, \mu \in \mathbb{R}$,

$$\Lambda(\lambda, \mu) \;=\; \log \mathbf{E}\left[e^{\lambda Z}\right] + \mu \log d \;=\; \log\left(\sum_{i=1}^{d} e^{i\lambda}\right) + (\mu - 1)\log d.$$

For $x \in [1, d]$, there exists $\lambda = \lambda(x)$ such that

$$x = \left.\frac{\partial \Lambda(\lambda, \mu)}{\partial \lambda}\right|_{(\lambda(x),1)} = \frac{\sum_{i=1}^{d} i e^{i\lambda}}{\sum_{i=1}^{d} e^{i\lambda}}. \tag{6.37}$$

Then, we have

$$\Lambda^\star(x, y) = \begin{cases} \lambda x - \log\left(\sum_{i=1}^{d} e^{i\lambda}\right) + \log d & \text{if } x \in [1, d], y = \log d \\ \infty & \text{otherwise.} \end{cases}$$

As for ordinary tries, in the range of interest,

$$\phi\left(\alpha, \frac{1}{\log d}\right) = 1 - \alpha\lambda(\alpha \log d) + \frac{\Lambda(\lambda(\alpha \log d), 1)}{\log d}, \tag{6.38}$$

where $\lambda(\cdot)$ is defined in (6.37). In essence, $\phi(\alpha, t)$ is a function of $\alpha$ only. And we now write $\phi(\alpha) = \phi(\alpha, t)$ and $\Lambda(\lambda) = \Lambda(\lambda, 1)$. By Theorem 6.7, looking for the constant $c_b$ boils down to finding $\alpha_o$ such that

$$\left.\frac{\partial\phi(\alpha)}{\partial\alpha}\right|_{\alpha=\alpha_o} = \log Q(b+1) = -b\log d,$$

and for this $\alpha_o$, we have

$$c_b = \alpha_o + \frac{\phi(\alpha_o)}{\log d}. \tag{6.39}$$

In other words, we have

$$\begin{aligned}
\left.\frac{\partial\phi(\alpha)}{\partial\alpha}\right|_{\alpha_o} &= -\lambda(\alpha_o \log d) - \alpha \left.\frac{\partial\lambda(\alpha\log d)}{\partial\alpha}\right|_{\alpha_o} + \frac{1}{\log d}\cdot\left.\frac{\partial\Lambda(\lambda(\alpha\log d))}{\partial\alpha}\right|_{\alpha_o} \\
&= -\lambda(\alpha_o \log d) - \alpha \left.\frac{\partial\lambda(\alpha\log d)}{\partial\alpha}\right|_{\alpha_o} + \frac{1}{\log d}\cdot\left.\frac{\partial\Lambda(\lambda)}{\partial\lambda}\right|_{\lambda(\alpha_o\log d)}\cdot\left.\frac{\partial\lambda(\alpha\log d)}{\partial\alpha}\right|_{\alpha_o} \\
&= -\lambda(\alpha_o \log d),
\end{aligned}$$

by (6.37), and hence $\lambda(\alpha_o \log d) = b\log d$. Hence, by (6.38) and (6.39),

$$c_b = \frac{1}{b\log d} + \frac{\Lambda(b\log d)}{b\log^2 d}.$$

Observe that this characterizes fully $c_b$ and holds for any symmetric weighted trie. For our case of symmetric list-tries, we obtain

$$c_b = c_b(d) = \frac{\log\left(\sum_{i=1}^{d} d^{bi}\right)}{b\log^2 d} \sim \frac{d}{\log d},$$

for large $d$. Some numerical values can be found in Table 6.2.

### 6.6.4 Ternary search trees

In the ternary search trees introduced by Bentley and Sedgewick (1997), the implementation of a node uses a binary search tree. Hence, the cost of branching to a

| $b$ | 1 | 2 | 3 | 10 | 50 | 100 |
|---|---|---|---|---|---|---|
| $c_b(2)$ | 3.28661... | 2.67491... | 2.52441... | 2.44289... | 2.44215... | 2.44206... |
| $c_b(3)$ | 3.12515... | 2.86870... | 2.83088... | 2.82022... | 2.81969... | 2.81963... |
| $c_b(10)$ | 4.92852... | 4.90959... | 4.90850... | 4.90723... | 4.90680... | 4.90675... |

***Table 6.2:*** *Some numerical values of $c_b = c_b(d)$ characterizing the height of symmetric list-tries.*

character $i \in \mathcal{A}$ at a node $u$ is the depth of $i$ in the binary search built from the (non-uniform) random permutation $\sigma_u$. When the node $u$ is of type $\tau_u = (1, \dots, 1)$, the permutation $\sigma_u$ is distributed as the ordered list of first appearances of characters in an infinite string generated by the memoryless source with distribution $\{p_1, \dots, p_d\}$.

Let $Z_i$ be distributed as the depth of $i$ in the binary search tree built from $\sigma$. Then, $\mathcal{Z}$ is distributed as $(Z_1, \dots, Z_d)$ and $Z = Z_K$, where $K$ is uniform in $\{1, \dots, d\}$. When $u$ is a non-branching node, i.e., $\tau_u$ is a permutation of $(1, 0, \dots, 0)$, then the depth of the unique child is always one: $Z^s = 1$ almost surely. By Theorem 6.1, we obtain:

**Theorem 6.8.** *Let $H_{n,b}$ be the weighted height of a $b$-TST on $n$ sequences. Let $\sigma$ be a permutation of $\{1, \dots, d\}$ built by sampling with replacement from $\{1, \dots, d\}$ according to $p_1, \dots, p_d$. Let $Z$ be the depth of a random node in a binary search tree built from $\sigma$. Let*

$$c_b = \sup_{\alpha, t > 0} \left\{ \alpha + \frac{\phi(\alpha, t)}{-\log Q(b+1)} \right\},$$

*where $\phi(\alpha, t)$ is the logarithmic profile defined in (6.10). Then, $H_{n,b} \sim c_b \log n$ in probability, as $n \to \infty$.*

The random variable $Z$ is complicated to describe in other terms for general distributions $p_1, \dots, p_d$. Some parameters like the average value and the variance of $Z_i$, $1 \le i \le d$, have been studied by Clément et al. (1998, 2001) and Archibald and Clément (2006). For this case, describing $Z$ and $\phi(\alpha, t)$ in a way that would lead to $c_b$ seems way more difficult than for list-tries.

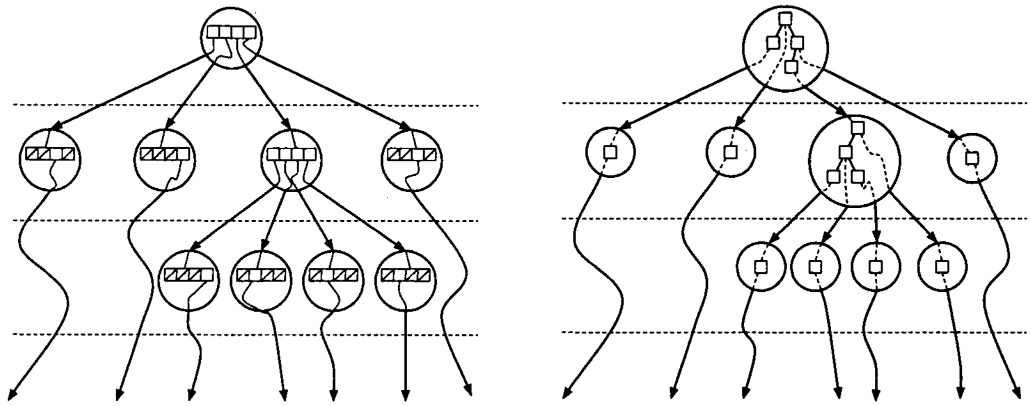

*Figure 6.11: A standart trie and the corresponding TST.*

**Example: Symmetric TST.** We assume here that $p_1 = p_2 = \cdots = p_d$. In this case, the permutation $\sigma$ is just a uniform random permutation. Hence, $Z_i$ is the depth of the key $i$ in a random binary search tree. Observe that unlike in the case of list-tries, $Z_i$, $1 \le i \le d$, do *not* have the same distribution. This is easily seen, since, for instance as $d \to \infty$, $\mathbf{E}Z_1 \sim \log d$ whereas $\mathbf{E}Z_{\lfloor d/2 \rfloor} \sim 2 \log d$. However, we are only interested in the distribution of $Z$, that is, the depth of a uniform random node. This distribution is known exactly for random binary search trees, and is due to Brown and Shubert (1984):

$$\mathbf{P}\{Z = k\} = \frac{2^{k-1}}{d \cdot d!} \sum_{j=k}^{d} \begin{bmatrix} d \\ j \end{bmatrix}, \tag{6.40}$$

where $\begin{bmatrix} n \\ k \end{bmatrix}$ denotes the Stirling number of the first kind with parameter $n$ and $k$, that is the number of ways to divide $n$ objects into $k$ nonempty cycles (see Sedgewick and Flajolet, 1996; Mahmoud, 1992b). Using (6.40), it is possible to compute the cumulant generating function $\Lambda$, and $\phi(\alpha, t)$. Numerical values for the constant $c_b = c_b(d)$ such that $H_n \sim c_b \log n$ in probability as $n \to \infty$ are given in Table 6.3. Observe that the when $d = 2$, TST are equivalent to list-tries. In general, using the computations we did in the case of symmetric list-tries,

$$\begin{aligned}
c = c(d) &= \frac{1}{\log d} + \frac{1}{\log^2 d} \log \left( \sum_{i=1}^{d} \sum_{j=i}^{d} \frac{2^{i-1}}{d \cdot d!} \begin{bmatrix} d \\ j \end{bmatrix} d^i \right) \\
&= \frac{1}{\log d} + \frac{1}{\log^2 d} \log \left( \frac{(2d) \cdot (2d+1) \cdots (3d-1) - d!}{d!(2d-1)} \right) \\
&\sim \frac{d(3 \log 3 - 2 \log 2)}{\log^2 d}
\end{aligned}$$

(see, e.g., Mahmoud, 1992b, p. 79). Numerical values for the constant $c = c(d)$ are given in Table 6.3.

| $b$ | 1 | 2 | 3 | 10 | 50 | 100 |
|---|---|---|---|---|---|---|
| $c_b(2)$ | 3.28661... | 2.67491... | 2.52441... | 2.44289... | 2.44215... | 2.44206... |
| $c_b(3)$ | 2.90777... | 2.66010... | 2.65121... | 2.65118... | 2.65117... | 2.65116... |

***Table 6.3:*** *Some numerical values of $c_b = c_b(d)$ characterizing the height of symmetric ternary search trees.*

*Chapter 7*

# Conclusion: Shedding light on trees

This thesis provides generalizations of the theorems about the asymptotic behavior of the heights of random trees. In Chapter 5, we showed

- that the study of heights of random trees benefits from the introduction of weighted versions of the standard branching process along the lines first suggested by Biggins (1977), and
- that, if there is an upper bound on the height, only the subtrees that contain a large number of items contribute significantly to the height.

This permits us to treat many kinds of trees using the same theorem (Theorem 5.1). Also, several new applications follow. In Chapter 6, we introduce an analogous weighted version of random tries. We proved that the height can be explained by

- the contribution of the core of the tree that behaves as weighted random split trees, and
- an additional term coming from long spaghetti-like trees.

These terms have comparable asymptotic growth, and in general, neither is negligible. The main result of Chapter 6, Theorem 6.1, allows us to devise new proofs for the heights of tries. New applications include the characterization of the asymptotic heights of the trees of de la Briandais (1959) and the ternary search tree of Bentley and Sedgewick (1997).

The analysis exhibits a strong link between random trees and random tries. The link lies in the core of the trees. We shall explain it using a concrete example. Consider $n$ sequences of characters generated by a memoryless source. The expected cores of the digital search tree (Chapter 5) and trie (Chapter 6) built from such sequences are similar. The trie differs in that the trees hanging down the core have height of order $\Theta(\log n)$. The heights of both trees may be explained by the *cone of shadow* cast by the logarithmic profile $\phi(\cdot, \cdot)$ describing the core. In the case of digital search tree, the bulb should be located far away in the direction $(0,0,1)$. In the case of tries, one should put the light at $(-\gamma_b, -\rho_b, 1) \cdot x$, for some specific positive $\gamma_b$ and $\rho_b$, and $x \to \infty$. So, in some sense, digital search trees and tries, appear as the same object seen from two different angles.

EXTENSIONS AND OPEN PROBLEMS. All the trees in the thesis have bounded degree, or are reduced to bounded degree. The theorems can be extended to the unbounded case, using the point process approach of Biggins (1995, 1996). Also, we only characterized the first order terms in the asymptotic expansion of the height. It would be interesting to see how general a theorem one can obtain about asymptotics that are precise to $O(1)$. We think in particular of the case of increasing trees (Bergeron et al., 1992; Broutin et al., 2006; Drmota, 2006) that may benefit from the new approach of Addario-Berry (2006) and Addario-Berry and Reed (2006).

# Bibliography

L. Addario-Berry. *Ballot Theorems and the Heights of Trees.* PhD thesis, McGill University, 2006.

L. Addario-Berry and B. Reed. personnal communication, 2006.

L. Addario-Berry and B. Reed. Ballot theorems, old and new. 2007.

N. Alon, J. Spencer, and P. Erdős. *The Probabilitic Method.* Wiley, New York, NY, second edition, 2000.

M. Archibald and J. Clément. Average depth in binary search tree with repeated keys. In *Fourth Colloquium on Mathematics and Computer Science*, 2006.

S. Arora and S. Safra. Probabilistic checking of proofs: a new characterization of NP. *Journal of the ACM*, 45:70–122, 1998.

K. B. Athreya and P. E. Ney. *Branching Processes.* Springer, Berlin, 1972.

M. Bachmann. Limit theorems for the minimal position in a branching random walk with independent logconcave displacements. *Advanced Applied Probability*, 32:159–176, 2000.

R. Baeza-Yates, R. Casas, J. Diaz, and C. Martinez. On the average size of the intersection of binary trees. *SIAM Journal on Computing*, 21(1):24–32, 1992.

R.R. Bahadur and R.R. Rao. On deviations of the sample mean. *Annals of Mathematical Statistics*, 31:1015–1027, 1960.

A.L. Barabási and R. Albert. Emergence of scaling in random network. *Science*, 286: 509–512, 1999.

J.-P. Barthélemy and A. Guénoche. *Trees and proximity representations.* Wiley, 1991.

J.-P. Barthélemy and N.X. Luong. Sur la topologie d'un arbre phylogénétique: aspects théoriques, algorithmiques et applications à l'analyse des données textuelles. *Math. Sci.*

*Hum.*, 100:57–80, 1987.

J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communication of the ACM*, 18:509–517, 1975.

J. L. Bentley and R. Sedgewick. Fast algorithm for sorting and searching strings. In *Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 360–369, 1997.

F. Bergeron, P. Flajolet, and B. Salvy. Varieties of increasing trees. In *CAAP*, volume 581 of *Lecture Notes in Computer Science*, pages 24–48, 1992.

P.K. Bhattacharya and J.L. Gastwirth. A nonhomogeneous markov model of a chain-letter scheme. In M.H. Rizvi, J.S. Rustagi, and D. Siegmund, editors, *Recent Advances in Statistics: Papers in Honor of Herman Chernoff.* Academic Press, New York, 1983.

J. D. Biggins. The first and last-birth problems in a multitype age-dependent branching process. *Advances in Applied Probability*, 8:446–459, 1976.

J. D. Biggins. Chernoff's theorem in the branching random walk. *Journal of Applied Probability*, 14:630–636, 1977.

J. D. Biggins. Spatial spread in branching processes. In *Biological Growth and Spread*, volume 38 of *Lecture Notesin Biomath.*, pages 57–67, Berlin, 1980. Springer.

J. D. Biggins. The growth and spread of the general branching random walk. *The Annals of Applied Probability*, 5:1008–1024, 1995.

J. D. Biggins. How fast does a general branching random walk spread. In K. B. Athreya and P. Jagers, editors, *Classical and modern branching processes*, New York, 1996. Springer-Verlag.

J. D. Biggins and D. R. Grey. A note on the growth of random trees. *Statistics and Probability letters*, 32:339–342, 1997.

P. Billingsley. *Probability and Measure.* Wiley, New York, 3rd edition, 1995.

M.D. Bramson. Maximum displacement of branching brownian motion. *Communications on Pure and Applied Mathematics*, 31:531–581, 1978.

M.D. Bramson and O. Zeitouni. Tightness for a family of recursive equations. Manuscript, 2006.

N. Broutin and L. Devroye. Large deviations for the weighted height of an extended class of trees. *Algorithmica*, 46:271–297, 2006.

N. Broutin and L. Devroye. The core of a trie. Manuscript, 2007a.

N. Broutin and L. Devroye. Weighted height of random tries. Manuscript, 2007b.

N. Broutin and L. Devroye. The height of list tries and TST. In *International Conference on Analysis of Algorithms*, 2007c. Accepted for publication.

N. Broutin, L. Devroye, E. McLeish, and M. de la Salle. The height of increasing trees. Submitted, 2006.

N. Broutin, L. Devroye, and E. McLeish. Weighted height of random trees. Manuscript, 2007.

G.G. Brown and B.O. Shubert. On random binary trees. *Mathematics of Operations Research*, 9:43–65, 1984.

B. Chauvin and M. Drmota. The random multisection problem, travelling waves, and the distribution of the height of $m$-ary search trees. *Algorithmica*, 2007. to appear.

H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

V. Choi and M. J. Golin. Lopsided trees I: a combinatorial analysis. *Algorithmica*, 31: 240–290, 2001.

C. Choppy, S. Kaplan, and M. Soria. Complexity analysis of term-rewriting systems. *Theoretical Computer Science*, 67:261–282, 1989.

H. A. Clampett. Randomized binary searching with tree structures. *Communications of the ACM*, 7(3):163–165, 1964.

J. Clément. *Arbres digitaux et sources dynamiques.* PhD thesis, Université de Caen, 2000.

J. Clément, P. Flajolet, and B. Vallée. The analysis of hybrid trie structures. In *9th annual ACM-SIAM Symposium on Discrete Algorithms*, pages 531–539, Philadelphia, PA, 1998. SIAM Press.

J. Clément, P. Flajolet, and B. Vallée. Dynamical source in information theory: a general analysis of trie structures. *Algorithmica*, 29:307–369, 2001.

E. G. Coffman and J. Eve. File structures using hashing functions. *Communications of the ACM*, 13:427–436, 1970.

H. Cohn. A martingale approach to supercritical (CMJ) branching processes. *The Annals of Probability*, 13(1179–1191), 1985.

T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction to algorithms.* MIT Press, Cambridge, MA, second edition, 2001.

H. Cramér. Sur un nouveau théorème-limite de la théorie des probabilités. In *Colloque Consacré à la Théorie des Probabilités*, volume 736, pages 5–23. Hermann, Paris, 1938.

R. de la Briandais. File searching using variable length keys. In *Proceedings of the Western Joint Computer Conference, Montvale, NJ, USA*. AFIPS Press, 1959.

A. Dembo and O. Zeitouni. *Large Deviation Techniques and Applications.* Jones and Bartlett, 1992.

A. Dembo and O. Zeitouni. *Large Deviation Techniques and Applications.* Springer Verlag, second edition, 1998.

F. den Hollander. *Large Deviations.* American Mathematical Society, Providence, RI, 2000.

J.-D. Deuschel and D.W. Stroock. *Large Deviations.* American Mathematical Society, Providence, RI, 1989.

L. Devroye. Laws of large numbers and tail inequalities for random tries and PATRICIA trees. *Journal of Computational and Applied Mathematics*, 142:27–37, 2002.

L. Devroye. Universal asymptotics for random tries and PATRICIA trees. *Algorithmica*, 42: 11–29, 2005.

L. Devroye. A probabilistic analysis of the height of tries and of the complexity of triesort. *Acta Informatica*, 21:229–237, 1984.

L. Devroye. A note on the height of binary search trees. *Journal of the ACM*, 33:489–498, 1986.

L. Devroye. Branching processes in the analysis of the heights of trees. *Acta Informatica*, 24:277–298, 1987.

L. Devroye. On the height of $m$-ary search trees. *Random Structures and Algorithms*, 1: 191–203, 1990.

L. Devroye. On the expected height of fringe balanced trees. *Acta Informatica*, 30:459–466, 1993.

L. Devroye. Branching processes and their application in the analysis of tree structures and tree algorithms. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, volume 16 of *Springer Series on Algorithms and Combinatorics*, pages 249–314, Berlin, 1998a. Springer.

L. Devroye. Universal limit laws for depth in random trees. *SIAM Journal on Computing*, 28(2):409–432, 1998b.

L. Devroye and B. Reed. On the variance of the height of binary search trees. *SIAM Journal on Computing*, 24:1157–1162, 1995.

L. Devroye, W. Szpankowski, and B. Rais. A note on the height of suffix trees. *SIAM Journal on Computing*, 21:48–53, 1992.

L. Devroye, J. Jabbour, and C. Zamora-Cura. Squarish $k$-d trees. *SIAM Journal on Computing*, 30:1678–1700, 2001.

M. Drmota. An analytic approach to the height of binary search trees. *Algorithmica*, 29: 89–119, 2001.

M. Drmota. An analytic approach to the height of binary search trees II. *Journal of the ACM*, 50:333–374, 2003.

M. Drmota. The height of increasing trees. *Annals of Combinatorics*, 2006. Submitted.

A. Duch. *Design and Analysis of Multidimensional Data Structures.* PhD thesis, UPC, Barcelona, 2004.

A. Duch and C. Martínez. On the average performance of orthogonal range search in multidimensional data structures. *Journal of the Algorithms*, 44(1):226–245, 2002.

R. Durrett. Maxima of branching random walks. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 62:165–170, 1983.

R. S. Ellis. Large deviations for a general class of random vectors. *The Annals of Probability*, 12:1–12, 1984.

R. S. Ellis. *Entropy, Large Deviations, and Statistical Mechanics.* Springer, New York, 1985.

W. Feller. *An Introduction to Probability Theory and its Applications*, volume I. Wiley, New York, 3rd edition, 1968.

W. Feller. *An Introduction to Probability Theory and its Applications*, volume II. Wiley, New York, 3rd edition, 1971.

P. Flajolet. Counting by coin tossings. In M. Maher, editor, *Proceedings of ASIAN'04 (Ninth Asian Computing Science Conference)*, volume 3321 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2004.

P. Flajolet. The ubiquitous digital tree. In B. Durand and W. Thomas, editors, *STACS 2006, Annual Symposium on Theoretical Aspects of Computer Science*, volume 3884 of *Lecture Notes in Computer Science*, pages 1–22, Marseille, February 2006.

P. Flajolet. On the performance evaluation of extendible hashing and trie searching. *Acta Informatica*, 20:345–369, 1983.

P. Flajolet and A. Odlyzko. The average height of binary trees and other simple trees. *Journal of Computer and System Sciences*, 25(171–213), 1982.

P. Flajolet and C. Puech. Partial match retrieval of multidimensional data. *Jounal of the ACM*, 33(2):371–407, 1986.

P. Flajolet and J.M. Steyaert. A branching process arising in dynamic hashing, trie searching and polynomial factorization. In M. Nielsen and E.M. Schmidt, editors, *Automata, Languages and Programming: Proceedings of the 9th ICALP Conference*, volume 140 of *Lecture Notes in Computer Science*, pages 239–251. Springer, 1982.

E. Fredkin. Trie memory. *Communications of the ACM*, 3(9):490–499, 1960.

M.L. Fredman and R.E. Tarjan. Fibonacci heaps and their uses in improved network optimization algorithms. *Jounal of the ACM*, 34:596–615, 1987.

F. Galton. Problem 4001. *Educational Times*, page 17, 1 April 1873.

F. Galton and H. W. Watson. On the probability of extinction of families. *J. Roy. Anthropol. Inst.*, 4:138–144, 1874.

M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* Freeman, San Francisco, 1979.

J. Gärtner. On large deviations from the invariant measure. *Theory of Probabability and its Applications*, 22:24–39, 1977.

J.L. Gastwirth and P.K. Bhattacharya. Two probability models of pyramid or chain letter schemes demonstrating that their promotional claims are unreliable. *Operations Research*, 32(3):527–536, 1984.

G. H. Gonnet and R. Baeza-Yates. *Handbook of Algorithms and Data Structures.* Addison-Wesley, Workingham, second edition, 1991.

G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes.* Oxford University Press, Oxford, second edition, 2001.

P. Groeneboom, J. Oosterhoff, and F.H. Ruymgaart. Large deviation theorems for empirical probability measures. *The Annals of Probability*, 7:553–586, 1979.

J. B. S. Haldane. A mathematical theory of natural and artificial selection, V. *Proceedings of the Cambridge Philosophical Society*, 23:838–844, 1927.

J. M. Hammersley. Postulates for subadditive processes. *The Annals of Probability*, 2: 652–680, 1974.

T. E. Harris. *The Theory of Branching Processes.* Springer, Berlin, 1963.

C. A. R. Hoare. Algorithm 63 and 64. *Communications of the ACM*, 4:321, 1961.

C. A. R. Hoare. Quicksort. *The Computer Journal*, 5:10–15, 1962.

Y. Hu and Z. Shi. Minimal position and critical martingale convergence in branching random walks, and directed polymers on disordered trees. URL `math.PR/0702799`. Manuscript, 2007.

D.A. Huffman. A method for constructing minimum redundancy codes. *Proceedings of the Institude of Radio Engineers*, pages 1098–1102, 1952.

H.K. Hwang. Local limit theorems for the profile of random tries. Manuscript, 2006.

P. Jacquet and M. Régnier. Trie partitioning process: limiting distributions. In *Proceedings of the 11th colloquium on trees in algebra and programming*, volume 214 of *Lecture Notes in Computer Science*, pages 196–210, New York, 1986. Springer.

P. Jagers. *Branching Processes with Biological Applications.* Wiley, New York, 1975.

S. Janson, T. Łuczak, and A. Ruciński. *Random Graphs.* Wiley, New York, 2000.

S. Kapoor and E. Reingold. Optimum lopsided binary trees. *Journal of the ACM*, 36(3): 573–590, july 1989.

D. G. Kendall. Branching processes since 1873. *Journal of the London Mathematical Society*, 41:385–406, 1966.

H. Kesten and B. P. Stigum. A limit theorem for multidimensional galton-watson processes. *Annals of Mathematical Statistics*, 37:1211–1233, 1966.

J.F.C. Kingman. The first birth problem for an age-dependent branching process. *The Annals of Probability*, 3:790–801, 1975.

D. E. Knuth. *The Art of Computer Programming: Fundamental algorithms*, volume 1. Addison-Wesley, 1973a.

D. E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison-Wesley, 1973b.

D. E. Knuth. *The Art of Computer Programming: Sorting and Searching*, volume 3. Addison-Wesley, Reading, MA, 1973c.

A.G. Konheim and D.J. Newman. A note on growing binary trees. *Discrete Mathematics*, 4:57–63, 1973.

H. Mahmoud. Distances in plane-oriented recursive trees. *Journal of Computational and Applied Mathematics*, 41:237–245, 1992a.

H. Mahmoud. *Evolution of Random Search Trees.* Wiley, New York, 1992b.

H. M. Mahmoud. A strong law for the height of random binary pyramids. *The Annals of Applied Probability*, 4:923–932, 1994.

C. Martínez and S. Roura. Optimal sampling strategies in quicksort and quickselect. *SIAM Journal on Computing*, 31:683–705, 2001.

C. Martínez, A. Panholzer, and H. Prodinger. Partial match in relaxed multidimensional search trees. *Algorithmica*, 29(1–2):181–204, 2001.

A. Meir and J.W. Moon. On the altitude of nodes in random trees. *Canadian Journal of Mathematics*, 30:997–1015, 1978.

D. R. Morrison. PATRICIA — Practical Algorithm To Retrieve Information Coded in Alphanumeric. *Journal of the ACM*, 15:514–534, 1968.

O. Nerman. On the convergence of a supercritical general (C-M-J) branching process. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57:365–395, 1981.

J. Pach and P.K. Agarwal. *Combinatorial Geometry.* Wiley, New York, 1995.

C. H. Papadamitriou. *Computational Complexity.* Addison-Wesley, 1993.

G. Park, H.K. Hwang, P. Nicodème, and W. Szpankowski. Profile of tries. manuscript, 2006.

V. V. Petrov. *Sums of Independent Random Variables.* Springer-Verlag, Berlin, 1975.

B. Pittel. On growing random binary trees. *Journal of Mathematical Analysis and its applications*, 103:461–480, 1984.

B. Pittel. Asymptotic growth of a class of random trees. *The Annals of Probability*, 13: 414–427, 1985.

B. Pittel. Paths in random digital tree: limiting distributions. *Advanced in Applied Probability*, 18:139–155, 1986.

B. Pittel. Note on the height of random recursive trees and $m$-ary search trees. *Random Structures and Algorithms*, 5:337–347, 1994.

S.T. Rachev and L. Rüschendorf. Probability metrics and recursive algorithms. *Advances in Applied Probability*, 27:770–799, 1995.

B. Reed. How tall is a tree? In *STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 479–483, New York, NY, USA, 2000. ACM Press.

B. Reed. The height of a random binary search tree. *Journal of the ACM*, 50:306–332, 2003.

M. Régnier. On the average height of trees in digital search and dynamic hashing. *Information Processing Letters*, 13:64–66, 1981.

J. M. Robson. The height of binary search trees. *The Australian Computer Journal*, 11: 1511–153, 1979.

J. M. Robson. The asymptotic behaviour of the height of binary search trees. *Aust. Comput. Sci. Commun.*, page 88, 1982.

R. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, NJ, 1970.

U. Rösler. A fixed point theorem for distributions. *Stochastic Processes and their Applications*, 37:195–214, 1992.

U. Rösler and L. Rüschendorf. The contraction method for recursive algorithms. *Algorith-*

*mica*, 29:3–33, 2001.

H. Samet. *Applications of Spatial Data Structures: Computer Graphics, Image Processing, and GIS.* Addison-Wesley, Reading, MA, 1990a.

H. Samet. *The Design and Analysis of Spatial Data Structures.* Addison-Wesley, Reading, MA, 1990b.

R. Sedgewick. *Quicksort.* Garland Publishing, New York, 1975.

R. Sedgewick and P. Flajolet. *An Introduction to the Analysis of Algorithm.* Addison-Wesley, 1996.

M. Sipser. *Introduction to the Theory of Computation.* Course Technology, second edition, 2005.

D.D. Sleator and R.E. Tarjan. Self-organizing binary search trees. *Jounal of the ACM*, 32 (652–686), 1985.

R. T. Smythe and H. M. Mahmoud. A survey of recursive trees. *Theoretical Probability and Mathematical Statististics*, 51:1–27, 1995.

J. F. Steffensen. On Sandsynlingheden for at Afkommet uddø r. *Matem. Tiddskr. B*, pages 19–23, 1930.

W. Szpankowski. *Average Case Analysis of Algorithms on Sequences.* Wiley, New York, 2001.

W. Szpankowski. On the height of digital trees and related problems. *Algorithmica*, 6: 256–277, 1991.

J. Szymański. On a non-uniform random recursive tree. In M. Karoński and Z. Palka, editors, *Random Graphs '85*, volume 33 of *Annals of Discrete Mathematics*, pages 297–306, Amsterdam, 1987. North Holland.

R.E. Tarjan. *Data Structures and Network Algorithms.* CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, 1983.

M. Van Emden. Increasing the efficiency of quicksort. *Communications of the ACM*, 13: 563–567, 1970.

B. F. Varn. Optimal variable length codes (arbitrary symbol costs and equal code word probabilities). *Informat. Contr.*, 19:289–301, 1971.

J. S. Vitter and P. Flajolet. Average-case analysis of algorithms and data structures. In J. van Leeuwen, editor, *Handbook of theoretical computer science*, volume A: Algorithms and Complexity, pages 431–524. MIT Press, Amsterdam, 1990.

J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transaction on Information Theory*, 23:337–343, 1977.

J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transaction on Information Theory*, 24:530–536, 1978.